THESIS

IMPACT OF ACTUAL AND SELF-PERCEIVED BODY TYPE ON VISUAL PERCEPTION OF DISTANCES

Submitted by

Matthew Branan

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Spring, 2015

Master's Committee:

Advisor: Phil Turk

Jessica Witt
Ann Hess

ABSTRACT

IMPACT OF ACTUAL AND SELF-PERCEIVED BODY TYPE ON VISUAL PERCEPTION OF DISTANCES

We investigate several questions regarding the proposition that physical body size and one's image of their own body type affect the ability to make accurate judgements of distances. Data collected include subjects' guesses of distances of four cones set 10, 15, 20, and 25 meters away and the weight, BMI, and self-perception of body image for each of 67 subjects. Interest lies in determining the covariates that are most important in explaining one's ability to accurately judge distances and whether weight or BMI is the better explainer among the physical body size predictors. We utilize linear mixed models to account for correlation among each subjects' own distance guesses and to allow for flexible modeling of subject-specific effects. Flexibility is further promoted through use of model averaging techniques to account for model selection uncertainty inherent in typical approaches in which an analyst selects only one model from which inferences are made. A generalization of the coefficient of determination from ordinary linear models is made to the linear mixed model setting (R^2_{LMM}) in order to provide an additional goodness measure for fixed effects and for individual fixed effects themselves.

Baseline differences among subjects' ability to accurately judge distances are so vast that extracting the importance of the fixed effects becomes difficult. It is found that body size is a significant predictor of subjects' ability to accurately judge distances but body image is not at the 0.05 significance level. We recommend choosing weight over BMI as a predictor of guessing behavior based on information criteria, model averaging, and the generalized R^2_{LMM}. Specifically, heavier individuals tend to guess more accurately.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Phil Turk for extending his time and effort to help me realize completion of this investigation. His continual dedication to advancing understanding, fostering my ability as a statistician, and his general encouragement has helped me to ceaselessly improve upon myself throughout the entire process.

In addition, many thanks go out to Dr. Jessica Witt and Dr. Ann Hess for taking the time and effort to serve on the master's committee. Also, I am grateful for Dr. Witt's implementing the study described and for her willingness to allow a graduate student to analyze her data.

I would also like to thank my wife, Sarah, whose endless love and seemingly Athenian strength have helped me to achieve potentials previously unknown to me and for serving as a model of lasting progress.

Finally, I would like to acknowledge the love from my ma and pa. Without their direction and caring hands, where would I be if not here?

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Background

1.1 Scientific Questions of Interest

A question surfacing in psychological investigations is the following: how does one's perception of reality differ based on one's perception of the self? Dr. Witt has pursued this question in several applications including determining if baseball players with higher batting averages tend to see the baseball as being larger than do other players [34]. Other, related studies carried out by Dr. Witt include determining if parkour athletes with more favorable views of their own abilities saw walls as being shorter than less confident athletes [30] and if subjects suffering chronic pain tended to perceive distances as being longer [33].

In this application, we attempt to investigate the question: does one's physical size or one's perception of such affect their ability to judge distances? Say that a subject views themselves as having a certain body type and they perceive a cone that is placed near to them as being far away. This could serve as evidence of perceived pessimism that associates with learned hopelessness for those folks who are uncomfortable in their own bodies. In a related manner, this investigation can be used to identify whether physical limits of one's body size affects their ability to judge distances. For example, we might expect folks who are heavier to be taller, and those taller subjects to be able to judge distances more accurately due to their favorable visual vantage point higher from the ground than their shorter counterparts.

In our present problem, we would like to investigate the relationship between a subject's accuracy at judging distances and their real and perceived body types. To construct her set of data, Dr. Witt collected 67 subjects. Subjects were shown four cones that were placed at 10, 15, 20, and 25 meters away and they were asked to judge the distances between themselves and each of the four cones, in turn. In our analysis, we took these guessed distances as our response vectors in the subsequent analyses.

In addition to the four responses, the additional measurements made on each subject included their weight in pounds (denoted by *W*, where necessary), body mass index (*BMI*), and slef-perceived body type (*I* for image). In order to assess this last measurement, Dr. Witt presented a placard displaying silhouettes of varying body types (skinny and tall, short and stout, tall and stout, etc.) and individuals were asked to indicate the body type they believed to describe them best. Lastly, we denote the physical distances being guessed *D*.

It is important to note that the inclusion of subjects was carried out on a voluntary basis and no randomization scheme was employed. Inherent to this effect, any conclusions that we reach in our analysis are not validly generalizable to an overarching population. Instead, they will only be directly applicable to the sampled subjects. However, if considered with a liberal mindset and a healthy amount of caution, the results could very well serve as preliminary conclusions about a population or sub-population of interest in future studies (e.g., CSU students or CSU psychology majors).

In accordance with her data collection, Dr. Witt has expressed interest in the following questions.

(I) Does actual body type (as measured by weight or BMI) or perceived body type affect one's ability to accurately judge distances?

(II) Is weight or BMI a better predictor of accurate distance guesses?

Our analysis of the sample data will primarily focus on the pursuit of answers to questions (I) and (II), while making sure that our analysis uses appropriate approaches and modeling techniques.

Confounding a straightforward analysis involving a multiple linear regression model is the fact that the four repeated measurements made on each subject are correlated amongst themselves, violating the crucial, canonical assumption of independent responses found among ordinary regression models. To cope with this complication, we first investigated the literature on repeated measures analysis.

1.2 Repeated Measures Analysis

One aspect of this study that we must take into account when delving into our data is that multiple measurements are made on each subject. Studies with this type of structure are typically called repeated measures studies and these have a particularly unique characteristic: responses on the same individual tend to be correlated amongst themselves.

Take a simple example. Suppose a researcher is measuring subjects' blood pressure before drinking coffee and again after drinking coffee, we expect for the two measurements to be related in some way because the the two measurements are made on the same individual and we expect that any particular individual has some baseline blood pressure about which any instantaneous measurement will simply be a deviation. In fact, if the correlation among responses made by the same subjects is ignored, we pay the price by making inference that is often too conservative (e.g., confidence intervals that are too wide) and end up wasting information embedded in our data's special structure.

To see this, consider the simple, paired, setup in which X_1 is the blood pressure of an individual before drinking coffee and X_2 is the blood pressure of the same individual after drinking coffee. Then, we can take a look at the variance of the difference between these two measurements.

$$\begin{aligned}\operatorname{Var}(X_1 - X_2) &= \operatorname{Var}(X_1) + \operatorname{Var}(X_2) - 2\operatorname{Cov}(X_1, X_2)\\ &= \operatorname{Var}(X_1) + \operatorname{Var}(X_2) - 2\sqrt{\operatorname{Var}(X_1)\operatorname{Var}(X_2)}\operatorname{Corr}(X_1, X_2)\\ &= \sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho_{12}\end{aligned}$$

Compare this to the situation in which we treat the two measurements as independent, as we might naïvely do if we were to apply standard methods to repeated measures data.

$$\begin{aligned}\operatorname{Var}(X_1 - X_2) &= \operatorname{Var}(X_1) + \operatorname{Var}(X_2) - 2\operatorname{Cov}(X_1, X_2)\\ &\overset{indep}{=} \operatorname{Var}(X_1) + \operatorname{Var}(X_2) - 0\\ &= \sigma_1^2 + \sigma_2^2\end{aligned}$$

Thus, if the pair of observations on the same subject are positively correlated such that $\rho_{12} > 0$ then we will see a decrease in the variance of the difference in the measurements if we do happen to account for their correlation. This yields narrower confidence intervals and more powerful tests when compared to the situation in which we ignore this aspect of the data. By this very simple example, we see value in accounting for correlation among measurements made on the same subject and an extension to more than 2 measurements per subject and to more than one subject results in similar revelations.

Historically, there have been several approaches to handling the correlation among subject's responses. These include models such as the repeated measures ANOVA, the use of summary measures, and the linear mixed model formulation. The repeated measures ANOVA took advantage of Fisher's analysis of variance approaches popularized in the early 20^{th} century. In this approach, the response vector of k response measurements for the i^{th} subject ($\mathbf{Y}_i$) is supposed to be a function of a matrix of categorical predictor variables ($\mathbf{X}_i$), a random, subject-specific variable (u_i), and random error ($\boldsymbol{\epsilon}_i$):

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{1}u_i + \boldsymbol{\epsilon}_i$$

where $i = 1, \ldots, n$ indexes subject. As is typical of ANOVA-type analyses, the researcher may then partition the various sources of variability and use overall F tests (or multiple comparisons procedures) to test for differences among groups of a treatment factor, among measurement occasions, or an interaction of the two.

The inherent problems associated with using a repeated measures ANOVA approach to repeated measures data include the following: (1) the restriction of having categorical predictor variables can hamper investigations in which continuous predictor variables are hypothesized to be related to the response, (2) it is difficult to manipulate the model to take into account messy data forms such as those with missing measurements or unequal time spacing between measurements, and (3) the model makes stringent assumptions about the covariance structure of the response vectors. Namely, it is assumed that the variance of the j^{th} measurement on the i^{th} subject, $\mathrm{Var}(Y_{ij}) = \sigma_u^2 + \sigma_\epsilon^2$ and the covariance between any two measurements on the same subject, $\mathrm{Cov}(Y_{ij}, Y_{ik}) = \sigma_u^2$ for $j \neq k$ – this covariance structure is called compound symmetry.

Verbeke and Molenberghs (1997) [32] describe the historic use of summary measures to reduce the dimensionality of the response vectors. In this approach, summary measurements are obtained by, for example, taking the mean of the vector of responses or by fitting a polynomial curve to the responses for a subject and calculating the area under this curve. The k responses for the n subjects are then reduced to one response per subject and ordinary regression, ANOVA, or other ordinary statistical analysis methods can be applied directly to the now one-dimensional responses. Although computationally attractive, a problem with this approach is that missing data or measurements made at unequal times between subjects is difficult to rectify. In addition, the summarization of the k responses with one value sacrifices information and eliminates the salient and often information-rich feature of these types of studies: the correlation among a subject's responses.

To avoid the problems posed above, we next look to a more flexible approach reminiscent of the well-documented linear model by examining the linear mixed model (LMM).

1.3 Linear Mixed Model Framework

Correlation amongst responses violates the ordinary multiple regression assumptions of independent observations. In order to cope with this, Fitzmaurice, Laird, and Ware [11] propose the use of a linear mixed model. In this model, assume that we collect k_i response measurements (across k_i occasions which could be across time, space, or even a simple series of measurements, but here we will simply call them occasions) on each of *n* subjects – note that we could have taken different numbers of measurements on each subject but our data are balanced and so we consider only the case in which $k_i = k$ for all $i = 1, \ldots, n$. We propose that the vector of responses from each subject is some linear combination of a set of *p* fixed predictor variables and *q* random predictor variables. That is,

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\epsilon}_i \tag{1}$$

Where

$\mathbf{Y}_i$: vector of k responses for subject i

$\mathbf{X}_i$: matrix of covariates for subject i (for p fixed effects)

$\boldsymbol{\beta}$: vector of fixed effects parameters

$\mathbf{Z}_i$: matrix of covariates for subject i (for q random effects)

$\mathbf{u}_i$: vector of random effects parameters for subject i

$\boldsymbol{\epsilon}_i$: vector of random errors for subject i

Traditionally, the random effects and the random errors are assumed mutually independent and normal

$$\mathbf{u}_i \overset{indep}{\sim} \mathrm{N}_q(\mathbf{0}, \mathbf{G})$$
$$\boldsymbol{\epsilon}_i \overset{indep}{\sim} \mathrm{N}_k(\mathbf{0}, \mathbf{R}_i)$$

where $\mathbf{G}$ is a covariance matrix for the random effects and $\mathbf{R}_i$ are covariance matrices for the random errors for subject i, each taking on user-specified structures. Furthermore, $\mathrm{Var}(\boldsymbol{\epsilon}) = \mathbf{R}$ is assumed to be a block-diagonal matrix with matrices $\mathbf{R}_1, \ldots, \mathbf{R}_n$ along the diagonal and zeros elsewhere to reflect the independence of random errors between subjects.

In this framework, the vectors of k response measurements are referred to as response profiles. The population is assumed to have a mean response profile that is averaged across all of the subjects, unconditional (or marginal) on subject-specific effects:

$$\mathbb{E}(\mathbf{Y}_i) = \mathbb{E}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\epsilon}_i)$$

$$
\begin{aligned}
&= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbb{E}(\mathbf{u}_i) + \mathbb{E}(\boldsymbol{\epsilon}_i) \\
&= \mathbf{X}_i\boldsymbol{\beta} \qquad (2)
\end{aligned}
$$

Not only can we average across the population, but we can also average across individual subject responses, conditional on subject-specific effects:

$$
\begin{aligned}
\mathbb{E}(\mathbf{Y}_i|\mathbf{u}_i) &= \mathbb{E}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\epsilon}_i|\mathbf{u}_i) \\
&= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbb{E}(\mathbf{u}_i|\mathbf{u}_i) + \mathbb{E}(\boldsymbol{\epsilon}_i|\mathbf{u}_i) \\
&= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i \qquad (3)
\end{aligned}
$$

which denotes the mean response profile for subject *i*. Each individual subject, then, is modeled as having their own mean response profile being a normal deviate about the population mean response profile and each individual response is a normal deviate about that subject's own mean profile. In this way, we can naturally model the between-subject variability and within-subject variability as a direct consequence of the model formulation.

Using these main ideas, we can employ a special case of the above linear mixed model called the random coefficients model. In this special case, the matrices $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ contain coefficients of best-fitting polynomials (lines, if each subject's matrix contains only the identity and another predictor) to the subject's responses. Within the special case, we can test for the population-wide fixed effects and their influence on the response as well as the effect of the uniqueness of subjects themselves and can visualize these via spaghetti plots. This idea is a simplification that offers itself to convenient, intuitive interpretations as one can imagine a population curve about which subject-specific curves vary randomly. Even though it is a special case intended to ease implementation by practitioners, the theoretical and computational conveniences inherent in the more general linear mixed model carry over directly.

1.4 Linear Mixed Model Estimation

With a model formulation in mind, the problem now becomes one of estimation; how might one estimate the components and the variances of the model in Equation (1). In our treatment, we will consider the model from a frequentist standpoint. In this way, we only attribute randomness to the error terms (which is propagated through to the responses) and the effects that we call random, the $\mathbf{u}_i$'s. A historically popular approach involves likelihood methods. Within the likelihood paradigm, there are two main methods of estimation in the linear mixed model: maximum likelihood (ML) and restricted maximum likelihood (REML). These two estimation procedures differ in their implementation and intended use and we visit each briefly.

In maximum likelihood estimation, we seek to maximize the likelihood of the unconditional responses with respect to both the regression coefficients, $\beta_0, \ldots, \beta_p$, and the variance terms,

$$\begin{aligned}
\mathrm{Var}(\mathbf{Y}_i) &= \mathrm{Var}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\epsilon}_i) \\
&= \mathrm{Var}(\mathbf{Z}_i\mathbf{u}_i) + \mathrm{Var}(\boldsymbol{\epsilon}_i) \\
&= \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}_i \qquad (4) \\
&\equiv \mathbf{V}_i
\end{aligned}$$

simultaneously. That is, we maximize:

$$\begin{aligned}
L(\boldsymbol{\beta}, \mathbf{V}_i|\mathbf{Y}_i) &= (2\pi)^{-k/2} \cdot |\mathbf{V}_i|^{-1/2} \cdot \exp\left\{-\frac{1}{2}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})^T\mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})\right\} \\
\Longrightarrow\ L(\boldsymbol{\beta}, \mathbf{V}|\mathbf{Y}) &= (2\pi)^{-nk/2} \cdot \prod_{i=1}^{n}|\mathbf{V}_i|^{-1/2} \cdot \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})^T\mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})\right\} \\
\Longrightarrow\ \ell(\boldsymbol{\beta}, \mathbf{V}|\mathbf{Y}) &= -\frac{nk}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\log|\mathbf{V}_i| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})^T\mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})
\end{aligned}$$

where $L(\cdot)$ and $\ell(\cdot)$ denote likelihoods and log likelihoods, respectively, $\mathbf{Y}_1, \ldots, \mathbf{Y}_k$ are concatenated column-wise to construct $\mathbf{Y}$, $\mathbf{X}_1, \ldots, \mathbf{X}_k$ are concatenated column-wise to construct $\mathbf{X}$, and $\mathbf{V}_1, \ldots, \mathbf{V}_k$ are aligned in a block-diagonal matrix $\mathbf{V}$.

The maximization leads to estimators $\widehat{\boldsymbol{\beta}}_{ML}$ and $\widehat{\mathbf{V}}_{ML}$. Note that the estimator for the regression coefficients is derived similarly and take the same form as weighted least squares estimates [18]. In general, we have,

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{ML} &= \left(\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-}\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{Y} && (5)\\
\mathrm{Var}(\widehat{\boldsymbol{\beta}}_{ML}) &= \mathrm{Var}\left(\left(\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-}\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{Y}\right)\\
&= \left[\left(\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-}\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\right]\hat{\mathbf{V}}\left[\widehat{\mathbf{V}}^{-1}\mathbf{X}\left(\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-}\right]\\
&= \left(\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-}\left(\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{X}\right)\left(\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-}\\
&= \left(\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-} && (6)
\end{aligned}$$

The estimator $\widehat{\boldsymbol{\beta}}_{ML}$ takes a normal distribution with the above mean and variance assuming the model assumptions hold. The estimator derived from the same maximization for the covariance matrix contains the true regression coefficients, $\boldsymbol{\beta}$, and one typically uses 'hat estimation' by plugging in the maximum likelihood estimates into the equation during computation. Estimation such as this does not take into account the extra uncertainty inherent in using an estimator in the place of true parameter values. This is one of the several disadvantages of the ML estimator for the variance components. Another disadvantage is the fact that the elements of $\widehat{\mathbf{V}}$ are typically biased, especially when the number of regressors, p, is near the overall sample size, nk [15].

In order to overcome these, while retaining the advantages of maximum likelihood estimators of asymptotic normality and efficiency, and hence consistency [7], another likelihood-based method is often used to derive an estimator for the covariance components of the model, REML. The goal behind REML estimation is to achieve an unbiased estimator for $\mathbf{V}$ by transforming the data so that the distribution of the transformed data is independent of the regression coefficients, $\boldsymbol{\beta}$. Harville (1977) demonstrates how to perform this transformation by taking linear combinations of the responses by multiplying $\mathbf{K}^T\mathbf{Y}$ such that $\mathbf{X}^T\mathbf{K} = \mathbf{0}$. Rao (1962) [25] showed that the general solution to this system of equations takes the form,

$$
\begin{aligned}
\mathbf{K} &= (\mathbf{I} - (\mathbf{X}^T)^- \mathbf{X}^T)\mathbf{A} \\
&= (\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^- \mathbf{X}^T)\mathbf{A} \\
&= (\mathbf{I} - \mathbf{H})\mathbf{A} \\
\implies \mathbf{K}^T\mathbf{Y} &= \mathbf{A}^T(\mathbf{I} - \mathbf{H})\mathbf{Y} \\
&= \mathbf{A}^T\mathbf{e} \qquad (7)
\end{aligned}
$$

where $\mathbf{X}^-$ denotes a generalized inverse of the matrix $\mathbf{X}$, $\mathbf{A}$ is some arbitrary matrix, $\mathbf{H}$ denotes the so-called hat-matrix from linear regression, and $\mathbf{e}$ is a matrix of regression residuals. From Equation (7), we see that the transformed data are really a subset of the residuals, revealing the meaning of another form of the acronym REML: *residual* maximum likelihood. Also, we can see from both Equation (7) and from the fact that the transformed responses now have a mean of zero (inherited from properties of the residuals) that the influence of the regression coefficients has been nullified by this transformation since the original responses' distribution only depended upon the regression coefficients through the mean response. The resulting likelihood for the new data was shown by Harville (1974) [14] to be:

$$
\begin{aligned}
L(\boldsymbol{\beta}, \mathbf{V}|\mathbf{K}^T\mathbf{Y}) &= (2\pi)^{-(n-p)/2} \cdot \prod_{i=1}^{n} |\mathbf{V}_i|^{-1/2} \cdot \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})^T\mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})\right\} \\
&\qquad \cdot \left|\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{X}_i\right|^{1/2} \cdot \left|\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i\right|^{-1/2} \\
\implies \ell(\boldsymbol{\beta}, \mathbf{V}|\mathbf{K}^T\mathbf{Y}) &= -\frac{n-p}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\log|\mathbf{V}_i| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})^T\mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}) \\
&\qquad + \frac{1}{2}\log\left|\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{X}_i\right| - \frac{1}{2}\log\left|\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i\right|
\end{aligned}
$$

The above likelihood can be maximized to find $\widehat{\mathbf{V}}_{REML}$, which will no longer have the disadvantage of being biased as under ML estimation. Maximization of the log likelihood is not trivial and is often done numerically.

We next turn briefly to the problem of 'estimating' the random coefficients, $\mathbf{u}_i$ for $i = 1, \ldots, k$. Instead of observing this estimation as an ordinary estimation problem, we can instead think of it as a prediction problem. Referencing Equation (3), we see that the subject-specific portion of the LMM, $\mathbf{Z}_i\mathbf{u}_i$ is simply the additive effect of that particular subject's random effect on the overall mean response. Thus, estimating $\mathbf{u}_i$ is akin to predicting the portion of the i^{th} subject's mean response profile that is attributable to that subject and not the overall population.

Taking from ordinary, Gaussian multivariate linear models [17], we know that we can construct the best linear unbiased predictor (BLUP) for the subject-specific random effects with methods similar to predicting the expected value of a set of future responses based on those we have already observed. If we consider the column-wise concatenation of the response vectors and random coefficients for all k subjects following a bivariate normal distribution as elicited by the ordinary LMM assumptions,

$$\begin{bmatrix} \mathbf{Y} \\ \mathbf{u} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R} & \mathbf{Z}\mathbf{G} \\ \mathbf{G}\mathbf{Z}^T & \mathbf{G} \end{bmatrix} \right)$$

we can obtain the BLUP for the random effects by estimating the conditional mean of the random effects given the observed data. This estimation follows directly from the construction of the conditional distribution of $\mathbf{u}|\mathbf{Y}$.

$$\mathbf{u}|\mathbf{Y} \sim N\left(\mathbf{G}\mathbf{Z}^T\left(\mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}\right)^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}), \mathbf{G} - \mathbf{G}\mathbf{Z}^T\left(\mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}\right)^{-1}\mathbf{Z}\mathbf{G}\right)$$

Therefore, if we employ 'hat estimation' by substituting the estimated fixed effects in for their true counterparts, then the estimated mean random effects given the observed data can be expressed as,

$$\widehat{\mathbf{u}} = \mathbb{E}(\mathbf{u}|\mathbf{Y}) = \widehat{\mathbf{G}}\mathbf{Z}^T \left(\mathbf{Z}\widehat{\mathbf{G}}\mathbf{Z}^T + \widehat{\mathbf{R}}\right)^{-1} \left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \tag{8}$$

which is referred to as the empirical best linear unbiased predictor (EBLUP) due to our using the estimated vector of fixed effects parameters and estimated covariance matrices using the user's estimation method of choice.

Now that we have estimation procedures for all three major components of the LMM, we can return to the issue of identifying the proper estimation procedure. There are two main situations in which we will use our estimation methods: in choosing among models and estimating both fixed effects ($\boldsymbol{\beta}$ terms) and covariance terms (parameters composing $\mathbf{V}$).

Consider the problem of choosing among competing models first. At the behest of many authors [11, 20, 21, 32, 35] one should use ML to compare models with the same covariance structure but differing fixed effects. When choosing among models with differing covariance structures but the same fixed effects one should use REML-based likelihoods. The reasons for the heavy hand are many.

First, even if we are comparing non-nested mean models with information criteria, one must be aware of differences among the ML-based and REML-based information criteria as specified by the software used by the analyst. For example, in SAS 9.4 [26] the information criteria under ML penalize on $p + q^*$ (the total number of fixed effects *and* covariance parameters estimated in the model) but the REML model size penalty is q^* (the number of covariance parameters estimated). Thus, REML comparisons with information criteria don't even take into account the number of fixed effects, further demonstrating the invariance of REML estimates to changes in the model for the mean and the need for comparing mean structures via ML estimation. We prefer REML for comparing covariance structures, given a set of fixed effects, since we yield the advantages discussed above.

When we turn to the problem of estimation, we must take into account that in this application, we intend to perform model averaging. We plan to use weights that depend on the type of estimation procedure being utilized. We plan to rank models using criteria derived from ML estimation. If we were to use REML to construct model averaging weights, this might be considered improper because if we compare models with different fixed effects using this procedure, we are comparing models fit to completely different sets of data – this stems from the fact that in REML, the transformation of the responses to get rid of the dependence on the fixed effects relies on the fixed effects whose influence we are eliminating. Therefore, to retain the original set of data and maintain a cohesive picture of the information contained in those data, we will use ML estimation to construct our weights.

What method should we now use to estimate the models to be ranked by such weights? We could use REML to avoid the bias imposed on the covariance estimators, but again, this would be disingenuous because we have constructed weights based on ML techniques in order to avoid comparing models based on entirely different transformations of the data. Therefore, the most appropriate estimation method left to us is ML. Even though this produces biased covariance parameter estimates, this can be defended by our moderate sample size and the opinion that we should remain consistent in our approach.

We do note, however, that there has been some opposition to the staunch stance against using REML-based methods to choose among models with differing fixed effects. Notably, Gurka (2006) [13] recently performed several simulation studies in which he attempted to select what he knew to be a true underlying model using ML and REML-based information criteria. In his investigation, he varied only the true mean model in some simulations and varied both the true mean and covariance structures in others and showed that the REML information criteria were able to choose the true model for the mean a comparable proportion of the time as ML-based methods. Even with this enlightening counter-example, however, we stick with the suggestions discussed above due to conventional reasoning and because the situation described by Gurka has not been investigated in more general settings as of yet.

1.5 Linear Mixed Model Diagnostics

In order to assess the fit of the models, we appeal to standard residual diagnostics. Historically, residual diagnostics have been well documented in the case of the ordinary linear model, but the literature is more sparse in terms of diagnostic tools applicable to the linear mixed model specifically. As a result, many of the diagnostic tools have been adopted from the ordinary linear regression and the multiple linear regression settings.

One of the features of the linear mixed model that complicates the diagnosis adequate model fit issues is that there are the two separate, but related, structures for the covariance and the mean that make up the model as a whole. In the same thread, there are two separate, but related, formulations of what we might call residuals of the LMM. These directly relate to the between-subject (population averaged) and within-subject (subject-specific) effects from Equations (2) and (3). That is, we can have either of,

$$\begin{aligned} \mathbf{e}_{i,marg} &= \mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}} && \text{(marginal)} \\ \mathbf{e}_{i,cond} &= \mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}} - \mathbf{Z}_i\widehat{\mathbf{u}}_i && \text{(conditional)} \end{aligned}$$

The marginal residuals measure the deviations of the observed responses about the overall mean response profile while the conditional residuals measure the deviations of the observed responses for specific subjects about that subject's own mean response profile. These two types of residuals measure slightly different aspects of the model. Firstly, both types can be used to judge the adequacy of the mean structure. The former measure the ability of the chosen between-subject (i.e., fixed) effects to describe the data and the latter are aimed at measuring how well the choices of both between- and within-subject (i.e., fixed and random) effects model the mean structure of the data. Secondly, the marginal residuals focus more on the contribution of the within-subjects covariance matrices $\mathbf{R}_1, \ldots, \mathbf{R}_n$ while the conditional residuals can be used to pinpoint the consequences of imposing certain forms on the entire response covariance structure induced by $\mathbf{V}$. That is, both can be used to diagnose whether these covariance structures adequately model the sources of variance in the data (subject-specific variance across occasion and the variance of he random coefficients) [27].

Of course, just as in ordinary regression settings, there are problems with using the raw residuals. The most egregious characteristic of the raw residuals stems from the fact that they are but predictions of the true error terms. As such, they are random variables and typically inherit heteroskedasticity from the fact that predictions made further from the bulk of the data are more imprecise. As such, we might spuriously declare a point far from the bulk of the data, in the directions of the covariates, an outlier when, really, we expect points far out in the directions of the covariates to vary much more than those closer to the bulk of the data. To mitigate this characteristic of the raw residuals, one may standardize them according to their estimated variances – which are functions of the covariates. This way, we may more accurately judge points as being outliers. We accomplish this by first deriving the estimated variances of the marginal and conditional residuals. For the following, let $\mathbf{P}_i = \mathbf{X}_i \left(\mathbf{X}_i^T \widehat{\mathbf{V}}_i^{-1} \mathbf{X}_i \right) \mathbf{X}_i^T$. Notice that $\mathbf{P}_i = \mathbf{P}_i^T$ and,

$$\begin{aligned}
\mathbf{P}_i \widehat{\mathbf{V}}_i^{-1} \mathbf{P}_i &= \mathbf{X}_i \left(\mathbf{X}_i^T \widehat{\mathbf{V}}_i^{-1} \mathbf{X}_i \right) \left(\mathbf{X}_i^T \widehat{\mathbf{V}}_i^{-1} \mathbf{X}_i \right) \left(\mathbf{X}_i^T \widehat{\mathbf{V}}_i^{-1} \mathbf{X}_i \right) \mathbf{X}_i^T \\
&= \mathbf{X}_i \left(\mathbf{X}_i^T \widehat{\mathbf{V}}_i^{-1} \mathbf{X}_i \right) \mathbf{X}_i^T \\
&= \mathbf{P}_i
\end{aligned}$$

Then, we can derive,

$$\begin{aligned}
\widehat{\text{Var}}(\mathbf{e}_{i,marg}) &= \widehat{\text{Var}} \left(\mathbf{Y}_i - \mathbf{X}_i \widehat{\boldsymbol{\beta}} \right) \\
&= \widehat{\text{Var}} \left[\left(\mathbf{I} - \mathbf{P}_i \widehat{\mathbf{V}}_i^{-1} \right) \mathbf{Y}_i \right] && \text{(from Equation (5))} \\
&= \left(\mathbf{I} - \mathbf{P}_i \widehat{\mathbf{V}}_i^{-1} \right) \widehat{\text{Var}}(\mathbf{Y}_i) \left(\mathbf{I} - \mathbf{P}_i \widehat{\mathbf{V}}_i^{-1} \right)^T \\
&= \left(\mathbf{I} - \mathbf{P}_i \widehat{\mathbf{V}}_i^{-1} \right) \widehat{\mathbf{V}}_i \left(\mathbf{I} - \widehat{\mathbf{V}}_i^{-1} \mathbf{P}_i \right) \\
&= \widehat{\mathbf{V}}_i - \mathbf{P}_i - \mathbf{P}_i + \mathbf{P}_i \widehat{\mathbf{V}}_i^{-1} \mathbf{P}_i
\end{aligned}$$

$$= \widehat{\mathbf{V}}_i - \mathbf{P}_i$$

Similarly, we can find the variance of the conditional residuals. For the following, let $\mathbf{Q}_i = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i^T\mathbf{V}_i^{-1}$.

$$\begin{aligned}
\widehat{\mathrm{Var}}(\mathbf{e}_{i,cond}) &= \widehat{\mathrm{Var}}\left(\mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}} - \mathbf{Z}_i\widehat{\mathbf{u}}_i\right) \\
&= \widehat{\mathrm{Var}}\left[\left(\mathbf{I} - \mathbf{P}_i\widehat{\mathbf{V}}_i^{-1}\right)\mathbf{Y}_i - \left(\mathbf{Q}_i(\mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}})\right)\right] && \text{(from Equation (8))} \\
&= \widehat{\mathrm{Var}}\left[\left(\mathbf{I} - \mathbf{P}_i\widehat{\mathbf{V}}_i^{-1}\right)\mathbf{Y}_i - \left(\mathbf{Q}_i(\mathbf{I} - \mathbf{P}_i\widehat{\mathbf{V}}_i^{-1})\mathbf{Y}_i\right)\right] \\
&= \widehat{\mathrm{Var}}\left[(\mathbf{I} - \mathbf{Q}_i)\left(\mathbf{I} - \mathbf{P}_i\widehat{\mathbf{V}}_i^{-1}\right)\mathbf{Y}_i\right] \\
&= (\mathbf{I} - \mathbf{Q}_i)\left(\widehat{\mathbf{V}}_i - \mathbf{P}_i\right)(\mathbf{I} - \mathbf{Q}_i)^T
\end{aligned}$$

Therefore, to calculate the studentized marginal and conditional residuals, we can simply divide the raw residuals by their variances, as derived above. To do so, we can let, for example, $e_{ij,marg}$ be the j^{th} element of the i^{th} subject's marginal residual vector and $\widehat{\mathrm{Var}}(\mathbf{e}_{i,marg})_j$ denote the j^{th} diagonal element of the covariance matrix of the marginal residuals for the i^{th} subject. Our studentized residuals will then take the following form.

$$\begin{aligned}
e^s_{ij,marg} &= \frac{e_{ij,marg}}{\sqrt{\widehat{\mathrm{Var}}(\mathbf{e}_{i,marg})_j}} && \text{(studentized marginal)} \\
e^s_{ij,cond} &= \frac{e_{ij,cond}}{\sqrt{\widehat{\mathrm{Var}}(\mathbf{e}_{i,cond})_j}} && \text{(studentized conditional)}
\end{aligned}$$

Another problem with the raw residuals – specifically for the marginal residuals – from repeated measures problems is that they can be correlated and are typically heteroskedastic, a common trait seen in these types of data. This confounds the ability of some of the standard residual diagnostic tools one might ordinarily use. Therefore, it has been proposed [11, 26, 32] that one should scale the residuals by a Cholesky decomposition of the covariance matrix of the responses for each individual as shown below.

$$\widehat{\mathrm{Var}}(\mathbf{Y}_i) = \widehat{\mathbf{V}}_i$$
$$\widehat{\mathbf{V}}_i = \widehat{\mathbf{L}}_i\widehat{\mathbf{L}}_i^T$$

The second equality is the result of a Cholesky decomposition and $\mathbf{L}_i$ is a lower triangular matrix. Then, when the residuals are scaled by the inverse the Cholesky lower triangular matrix, $\mathbf{e_i^*} = \widehat{\mathbf{L}}_i^{-1}\mathbf{e}_i$, we will, theoretically, be left with residuals that are uncorrelated and have a standardized variance.

Once a choice of one or more types of residuals is made for use in a certain application, one can construct the usual residual diagnostic plots. A plot of the residuals versus the fitted values can be used to check for the lack of fit of both the mean and covariance structures simultaneously. If there is some sort of systematic trend in the plot then one should be aware of the possibility that there is likely some pattern inherent in the data that is not picked up by one or both components of the model. One may also check if there is heteroskedasticity, which might be indicative of an inappropriate covariance structure either in the specification of the forms of the two covariance matrices or in the random effects themselves. One should recall that we typically expect such behavior in the raw residuals in repeated measures data and so we might want to check this particular assumption using the scaled residuals instead.

Another canonical plot that may be used in the LMM context is the normal quantile-quantile (QQ) plot of the residuals. This plot is constructed by plotting the quantiles of the standardized residuals chosen by the user against the quantiles of the standard normal distribution. It allows one to observe the adherence of the specified model to the normality assumptions applied to the random effects and the errors. If there is evidence of deviation from a straight line in this plot then there is likely to be a violation of the normality assumptions, which indicates a misspecification of the distributional assumptions of the random components of the model.

The last characteristic one should investigate in the repeated measures setting is whether there are outlying observations. Again, just like there are two types of residuals that one may construct, there are also two types of outliers to check. The first are outliers with respect to the population averaged response profile. These include subjects whose entire mean response profile deviates from the response profile that has been averaged across all subjects. The second type of outlier include individual responses by specific subjects which deviate from that subject's mean response profile.

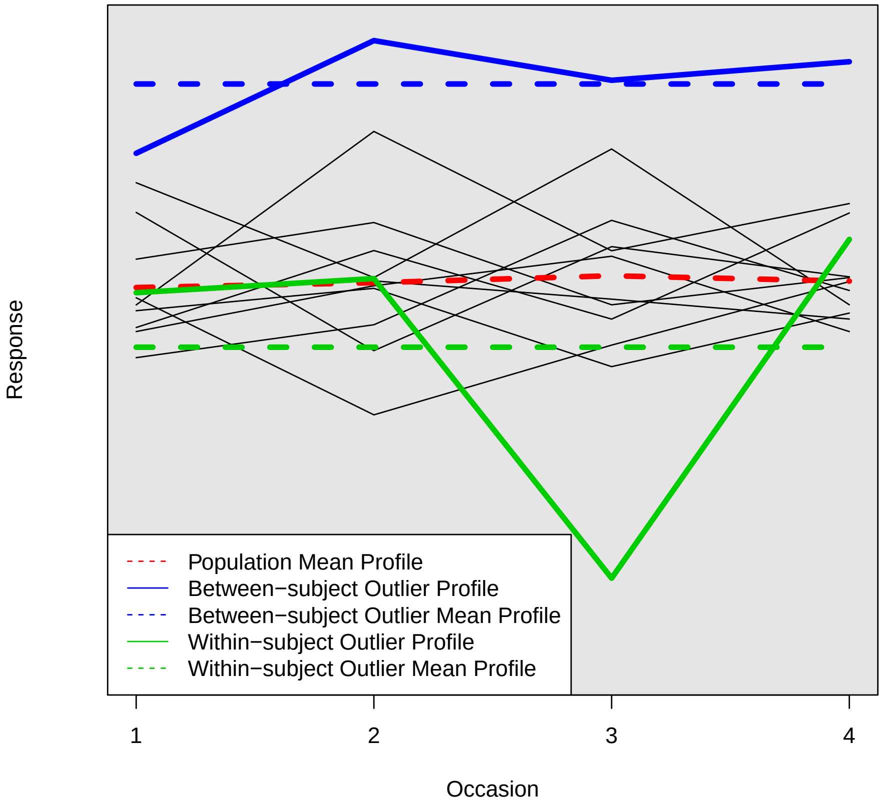


Figure 1.1: Illustration of the two different types of outliers in our repeated measures setting. The blue lines illustrate the between-subject outlier's raw response profile and their mean response profile while the green lines illustrate the same for the within-subject outlier. The population averaged response profile is included for reference as the red line.

The two types of outliers are illustrated in Figure 1.1. We assume that we have four measurement occasions and we measure some response on each subject for the four occasions. The between-subject outlier's raw responses are connected by the solid blue line and the mean response profile for this subject is the dashed blue line. Notice that this subject's entire mean response profile exceeds the population averaged profile and this deviation is visually sufficient to judge that this subject's response behavior is likely different from the average at every measurement occasion. The within-subject outlier is made by the subject whose raw and mean response profiles are plotted in green. This subject's response at the third measurement occasion seems to deviate from their own (and the overall) average response while the rest of the occasion measurements seem to align well with the average. This indicates that we likely have a within-subject outlier at the third occasion for this subject and that their response behavior may be different at that occasion only.

Besides visually inspecting for outlying values in the two residual plots mentioned above, a more direct way to detect outlying values is to employ leave-one-out methods. In leave-one-out methods, statistics are calculated after deleting one observation or subject at a time. If an observation or a subject are very influential in the construction of the quantities of interest, then we will see statistics that greatly change in value in a plot of those leave-one-out statistics.

One such statistic is Cook's distance [18, 26]. Cook's distance measures the influence of an observation or subject on the parameter estimates from the model. If we let $\widehat{\boldsymbol{\beta}}$ denote the vector of estimated fixed effects using all of the data and $\widehat{\boldsymbol{\beta}}_{(-i)}$ denote the same estimated parameters using all of the data except the i^{th} observation (or subject) then we would compute Cook's distances as follows.

$$D_i = \frac{\left(\widehat{\boldsymbol{\beta}}^T - \widehat{\boldsymbol{\beta}}^T_{(-i)}\right)\left(\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-}\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-i)}\right)}{rank\left(\mathbf{X}\right)}$$

Where these are computed for $i = 1, \ldots, n$ if looking at raw observations and $i = 1, \ldots, k$ if looking at the influence of subjects. The similarity of the form of the Cook's distances to F-test statistics (see Equation (13) below) has prompted many to compare them to quantiles of an F distribution in order to judge the influence of observations on the estimated parameters. Such an F distribution typically has numerator and denominator degrees of freedom equal to $rank\left(\mathbf{X}\right)$ and either $n - rank\left(\mathbf{X}\right)$ (for observations) or $k - rank\left(\mathbf{X}\right)$ (for subjects), respectively. The percentile against which one compares can be chosen by the user but is canonically the 25^{th} to judge borderline influential or the 50^{th} to judge influential points. These choices of percentiles mean that outliers are those points which result in standardized, absolute changes in the estimated parameters upon their exclusion that occur less than 50% of the time among repeated experiments. This statement assumes that the Cook's distances really are distributed as F random variables. Since this distributional assumption is not exact, the lenient cutoff of 50% is used rather than a more strict cutoff of, say, 90%.

Using these diagnostic tools, one can determine whether the model being fit is appropriate or not. If the fit happens to be questionable, one may implement a number of remedial measures including changing the forms of the covariance matrices or changing the random effects included in the model. The mean structure can be similarly changed by including more or deleting existing fixed effects from the model. Also, one can change the distributional assumptions of either the random effects, the random errors, or both simultaneously. Once the model fit seems adequate, one may then proceed to model selection and inference.

2 Methods

2.1 Akaike's Information Criterion (AIC) and Model Averaging

In 1974, Hirotugu Akaike expanded on his proposal of an information criterion that married the two foundational frameworks of information theory and likelihood theory [1]. In his paper, Akaike was looking to develop a criterion that would allow quantitative selection amongst several competing probability distributions in modeling the true nature of some real phenomenon. His contention with the hypothesis testing framework solidified by Neyman and Pearson was that the true nature of reality does not conform to the false dichotomy imposed by a choice among one null and one alternative hypothesis. Akaike might have agreed with George Box in saying that ”all models are wrong, but some are useful” [4]. In other words, reality is more complex than the models that we propose and those models that we do put forth are only approximations for the true nature of reality – and, we hope, at least some of those approximations are representative enough to produce meaningful conclusions and decisions.

The basis of the Akaike's information criterion (AIC) is derived from information theory, signal processing, and the quantification of entropy via the Kullback-Leibler difference in information. Consider a true data-generating process, t, which we attempt to approximate with some probability model based on a set of parameters, $f(\mathrm{x}|\boldsymbol{\theta})$. The difference in the true amount of information contained in t and our own model, on average, can be expressed as,

$$I(t, f(\mathbf{x}|\boldsymbol{\theta})) = \int t(\mathbf{x}) \log(t(\mathbf{x}))dx - \int t(\mathbf{x}) \log(f(\mathbf{x}|\boldsymbol{\theta}))dx$$

This represents the amount of information contained in full reality, t, that is lost by our approximation of such with the probability model $f(\mathrm{x}|\boldsymbol{\theta})$. Since t is unobserved, we cannot compute this quantity directly, but it can be estimated up to a constant. By several asymptotic approximations and distributional asymptotic theory, we end up with an information criterion [1],

$$\text{AIC} = -2\ell(\boldsymbol{\theta}|\mathbf{x}) + 2k$$

One issue with the ordinary AIC is that it is based on asymptotic theory and hence requires, much like an appeal to the central limit theorem, a large sample size in order to retain any optimality qualities supported by likelihood theory. However, Burnham and Anderson have supported another approximation to the Kullback-Leibler distance measure by using a second-order bias correction (akin to the second-order delta method) in the case of small to modest sample sizes. Their measure is called the corrected AIC[1],

$$\begin{aligned}\text{AICc} &= -2\ell(\boldsymbol{\theta}|\mathbf{x}) + 2k + \frac{2k(k+1)}{n-k-1} \\ &= \text{AIC} + \frac{2k(k+1)}{n-k-1}\end{aligned}$$

Presently, AIC is used in a couple of distinct ways. In one way, the AIC is used as a static model selection criterion. Say that we use AICc for model selection. Each model in a set would be assigned a value of AICc and that model with the most attractive value (that is, the lowest) is chosen as the ’best model’ and is then used for inference procedures as if this model was divined before the study began, ignoring all other models. Other criteria can be used in similar contexts, such as the Bayesian information criterion (BIC), Takeuchi’s information criterion (TIC), deviance information criterion (DIC).

Like Akaike and Box before them, Burnham and Anderson support the view that reality is not embodied by one or any number of models but is instead only approximated by such [5]. Due to their belief, the two have proposed performing mulitmodel inference with the help of information criteria. Instead of ranking models among a set of candidates, choosing one out of the set, and basing all inference on that single model, one should somehow incorporate the information contained

[1] In the coming analysis, we chose to employ the small-sample corrected Akaike information criterion and so focus our discussion for the remainder of this report on AICc.

in an entire set of candidate models and base inference on that collection of models, thereby accounting for the uncertainty embedded in the model selection process and avoiding model selection bias.

One method to accomplish this is to average estimates across models. Model averaging is prevalent in the literature and is practiced in a wide variety of applications including frequentist linear mixed models by Chen et al. (2013) [8], Bayesian models by Hoeting et al. (1999) [16], and, even though it is not stated explicitly by the authors, in a genetic optimization algorithm by Zhu et al. (2006) [36]. In essence, model averaging is an exciting prospect because it accomplishes two important tasks. First, it gives one a method to account for the uncertainty induced by model selection. Traditionally, one would typically select one model and ignore the uncertainty introduced by the choosing of that one model among all others. Second, model averaging, in regression especially, allows for an investigator to identify key covariates in their model without forcefully excluding those deemed unimportant by a likelihood ratio test or a similar approach. That is, all of the covariates that were considered to be important scientifically before analysis are retained in the model averaging approach, avoiding the problem of choosing a parsimonious model that eliminates information contained in those variables entirely – referred to as model selection bias.

The idea behind model averaging is simple. Before an analysis is begun – or, preferably, in the planning stages of the study – a set of possible models that explain the phenomenon under observation is constructed. For example, if one were conducting a study on disease dynamics, investigators might list important covariates such as genetic factors, health and nutrition of the specimens, and environmental conditions. A sequence of possible models explaining dynamics of the disease would then be constructed by taking either all subsets of those variables (if feasible) or at least a representative subset that are supported by scientific or logical reasoning.

The models chosen are then estimated and, in the case of information criteria-based model averaging, ranked according to the AICc. Model averaged estimates are then constructed by taking weighted averages of parameter estimates. The weighting scheme is based on AICc and the weighted averages are typically constrained to be convex combinations of the parameter estimates.

In deriving weights to be used in model averaging, we follow the logic of Burnham and Anderson (2004) [5]. Let us suppose that it has been predetermined that there are A models under consideration denoted by $M_1, \ldots, M_A$. Using an estimation method (e.g., ML or REML) we estimate the parameters of the models and compute AICc values for each, AICc$_1, \ldots,$ AICc$_A$. We may denote the ranked set of AICc values from the least to the greatest by AICc$_{(1)}, \ldots,$ AICc$_{(A)}$. Then, we may derive model weights based on the information criterion by first considering the differences between each model's information value and that from the model with the lowest AICc,

$$\Delta\text{AICc}_i = \text{AICc}_i - \text{AICc}_{(1)},\ i = 1, \ldots, A$$

Recall that since these AICc values are estimates of expected Kullback-Leibler distances, taking the difference of AICc values eliminates the unobservable constant term, $\ell(f(\mathbf{x}|\boldsymbol{\theta}^*))$. These give us an idea of the amount of information that is given up when we use model M_i in inference as opposed to the model with AICc$_{(1)}$ and leaves us with quantities that are immediately comparable to one another. Then, we can find a likelihood ratio comparing model M_i and the best model, $M_{(1)}$, scaled by a factor related to the difference in the number of estimable parameters under each model by exponentiating,

$$\begin{aligned}
\eta_i &\equiv \exp\left\{-\frac{1}{2}\Delta\text{AICc}_i\right\} \\
&= \exp\left\{-\frac{1}{2}\left[-2\ell_i(\widehat{\boldsymbol{\theta}}|\mathbf{x}) + \frac{2nk_i}{n-k_i-1}\right] - \left[-2\ell_{(1)}(\widehat{\boldsymbol{\theta}}|\mathbf{x}) + \frac{2nk_{(1)}}{n-k_{(1)}-1}\right]\right\} \\
&= \frac{L_i(\widehat{\boldsymbol{\theta}}|\mathbf{x})}{L_{(1)}(\widehat{\boldsymbol{\theta}}|\mathbf{x})} \cdot \exp\left\{\frac{nk_{(1)}}{n-k_{(1)}-1} - \frac{nk_i}{n-k_i-1}\right\}
\end{aligned}$$

The quantity η_i can be seen as the likelihood of model M_i given the sample data and conditioned on the set of models under investigation – in particular, the model $M_{(1)}$ [6, 2]. To derive AICc weights, we can transform the model likelihoods so that they sum to one,

$$w_i \equiv \frac{\eta_i}{\sum_{i=1}^{A} \eta_i} \tag{9}$$

If interpreted in terms of model likelihoods given the data, w_i can be seen as the probability that model M_i is optimal by the Kullback-Leibler distance measure within the set of A models considered. The weights allow direct computation of model averaged estimates of our parameters and subsequent model averaged inference, utilizing information aggregated across all of the models. Say that we are interested in generating a model averaged estimate of a particular parameter that is estimated in a model, $\widehat{\theta}_j$. An intuitive model averaged estimator for θ_j is then,

$$\widehat{\theta}_j^{MA} = \sum_{i=1}^{A} w_i \widehat{\theta}_{ij} \tag{10}$$

The construction of the estimator now includes two sources of uncertainty. Within each model, there is uncertainty in estimating θ_{ij}, as we would encounter in a traditional analysis, but now there is quantifiable uncertainty in the model selection procedure. An estimator of the variance can be derived,

$$\widehat{\mathrm{Var}}(\widehat{\theta}_j^{MA}) = \left(\sum_{i=1}^{A} w_i \sqrt{\widehat{\mathrm{Var}}(\widehat{\theta}_{ij}) + (\widehat{\theta}_{ij} - \widehat{\theta}_j^{MA})^2} \right)^2 \tag{11}$$

A relation between this estimator and mean squares in the ANOVA framework is made by relating within-group variability to the term $\widehat{\mathrm{Var}}(\widehat{\theta}_{ij})$ and between-group variability to the term $(\widehat{\theta}_{ij} - \widehat{\theta}_j^{MA})^2$. The major addendum here is that we are weighting the variances by the estimated model information content.

The estimator in Equation (11) is commonly referred to as an unconditional estimator of the variance of a particular parameter. This is due to the fact that we are averaging across models and are thus performing estimation without conditioning on a single model alone. Contrast this with the traditional method of estimation in which the variance of a parameter estimate is estimated using some single best model, which depends upon the model selected and would produce inference that would be too liberal since it ignored the uncertainty inherent in choosing that best model.

One item that we have not addressed yet is the situation in which not every model includes the j^{th} effect. Two common approaches deemed *simple* and *full* model averaging by Symonds et al. (2011) [29]. In the simple approach, one would recompute the AICc weights for the models containing the j^{th} effect and average over those model estimates *only*. In full model averaging, every model that does not contain parameter θ_j would simply be assumed to have restricted its estimate to exactly zero. In our treatment in the analysis to come, we employ the full model averaging approach since the weights are not dependent on the inclusion of specific effects in the various models and thus the weights are more reflective of the entire picture uncertainty contained in model selection, and implementation is straightforward.

Multimodel inference concerning model parameters may be carried out by using the estimators presented above. For example, say that one wished to estimate the effect of the first covariate in a linear model. A point estimate for the change in the response while holding other covariates at their respective levels could be computed by $\widehat{\beta}_1^{MA} = \sum_{i=1}^{A} w_i \widehat{\beta}_{i1}$. A $(1-\alpha/2)100\%$ confidence interval could be constructed by using a Wald interval [7]:

$$\widehat{\beta}_1^{MA} \pm z_{(1-\alpha/2)} \cdot \sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_1^{MA})}$$

where the variance is computed in Equation (11) and $z_{(1-\alpha/2)}$ is the value in the standard normal distribution such that $P(Z \leqslant z_{(1-\alpha/2)}) = \alpha/2$. The use of a Wald interval is justified by noting that estimation is carried out under the likelihood framework and thus the sampling distributions of the estimated regression coefficients are asymptotically normal. Since the model averaged estimators are linear combinations of asymptotically normal estimators then $\widehat{\beta}_1^{MA}$ should have a normal sampling distribution assuming sufficiently large sample sizes. Conclusions from inference can then be carried out in the conventional manner.

2.2 R^2 FOR FIXED EFFECTS

We next consider a coefficient of determination (R^2) for linear mixed models. In the ordinary linear regression (OLR) context, R^2 is a natural statistic computed by considering the proportion of variability in the response that is explained by its linear relationship with the covariates. In this simple setting, the proportion is computed as the ratio between measurements of the variability in the response explained by the model (SSM) and the total variability in the response (SST) – or an equivalent ratio involving the sum of squared residuals (SSR).

$$R^2_{OLR} = \frac{SSM}{SST} = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

where y_i is the i^{th} observed response, $\overline{y}$ is the mean response, and $\widehat{y}_i$ is the i^{th} predicted response from the model. Another interpretation involves the comparison of the above value to the overall F-test for model fit by considering a ratio of R^2_{OLR} quantities,

$$F_{OLR} = \frac{\frac{SSM}{df_M}}{\frac{SSR}{df_R}} = \frac{\frac{SSM}{df_M \cdot SST}}{\frac{SSR}{df_R \cdot SST}} = \frac{\frac{R^2_{OLR}}{df_M}}{\frac{1-R^2_{OLR}}{df_R}}$$

where df_M and df_R are the degrees of freedom for the model and residuals, respectively. In this light, R^2_{OLR} gives the ratio of the deviations of the responses from their predicted values from the intercept-only, or null, model and from the model in under observation – $\overline{y}$ being the predicted value for every observation in the intercept-only model.

This value is often used as a goodness criterion for a particular set of covariates answering the practical question: does this set of covariates explain an adequate amount of variability in the response? Such a use is different from information criteria with which we are only provided a relative quantification of model goodness – that is, information criteria answer the question: does this model result in smaller information loss than some other model? When we use AICc values, we cannot blindly interpret the raw values alone since they natively include a model-independent constant. Instead, we can take differences or a ratio of logarithms of information criteria to mitigate the influence of the constant. Thus, AICc is only interpreted in the context of comparing two or more models to one another. Contrarily, the coefficient of determination is a more direct measure of goodness of fit of a particular model and can be interpreted free of comparison with another model, if the user so wishes. In OLR, R^2 is popularly reported as a summary measure of a given model and is a comfort to practitioners. The search for a similar statistic in the LMM framework is what motivates this section.

When we move to the linear mixed model framework, there is no immediately obvious analogue to a coefficient of determination. There are many differences in model formulations as we move from an ordinary linear model to the linear mixed model. Douglas Bates (2008) [3] warns that one should be wary when making a generalization such as this by carefully thinking about what features of the ordinary coefficient of determination that we would like to carry over to the linear mixed model.

If we are attempting to develop a proportion of variation in the response explained by the set of covariates chosen, how should we define a residual? Should we consider the *unconditional* residuals or should we use the *conditional* residuals? Similarly, should we consider comparing fixed effects using the coefficient of determination, comparing the random effects, or should we be comparing the entire set of covariates included in any given model? Speaking of which, what is the intercept-only model? We could include a fixed intercept only ($\mathbf{Y}_i = \mathbf{1}_i\beta + \boldsymbol{\epsilon}_i$), a random intercept only ($\mathbf{Y}_i = \mathbf{1}_i u_i + \boldsymbol{\epsilon}_i$), or both of them ($\mathbf{Y}_i = \mathbf{1}_i\beta + \mathbf{1}_i u_i + \boldsymbol{\epsilon}_i$).

A few of the approaches to generalize an R^2 statistic have included pseudo R^2 values proposed for general linear models (especially in the case of binary response regression). Several of these involve taking a function of the likelihood ratio between the model of interest and the intercept-only model while others involve functions of variance terms from the model under study and the null model. We focus on one of the more recent attempts at implementing an R^2-like statistic in a linear mixed model context.

Edwards et al. (2008) [9] propose an R^2 statistic extending that from OLR by appealing to the interpretation of R^2_{OLR} as being a function of the F-test statistic comparing the model at hand to a null model. The proposed statistic compares models in which fixed effects are varied among models containing the same covariance structure. Also, the null model specified is one which contains an intercept for the fixed effects and retains the same structure as the model of interest in the random effects. We are effectively conducting a test of the hypotheses regarding which is the better of the two models: $\mathbf{Y}_i = \mathbf{1}_i\beta + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\epsilon}_i$ or $\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\epsilon}_i$ for $i = 1, \ldots, n$.

A natural way to test the above hypotheses is through a multivariate version of the overall F-test from OLR. That is, we can test the set of hypotheses $H_0 : \beta_1 = \ldots = \beta_p = 0$ versus $H_a : \exists\beta_i \neq 0$ for i in $1, \ldots, p$. Hypotheses such as these are special cases of the general overall F-test which can be conducted by considering a full column rank matrix of contrasts, $\mathbf{C}$, with rank $c \equiv \text{rank}(\mathbf{C})$. We effectively assume that the fixed effects are equal to zero,

$$\begin{aligned} H_0 &: \mathbf{C}\boldsymbol{\beta} = \mathbf{0} \\ H_a &: \mathbf{C}\boldsymbol{\beta} \neq \mathbf{0} \end{aligned} \tag{12}$$

An F-test statistic can be constructed by taking the squared distances between the contrasts of the fixed effects and the zero vector, weighted by their variance – recall Equation (6) for the variance of the estimator of the fixed effects [21].

$$F_{LMM} = \frac{\widehat{\boldsymbol{\beta}}^T \mathbf{C}^T \left[\mathbf{C} \left(\mathbf{X}^T \widehat{\mathbf{V}}^{-1} \mathbf{X} \right)^{-} \mathbf{C}^T \right] \mathbf{C}\widehat{\boldsymbol{\beta}}}{c} \tag{13}$$

Just like in the OLR case, one may assume that the above F-statistic is a function of some R^2 value in the same manner presented above. Thus, we should equate and solve,

$$\begin{aligned} F_{LMM} &= \frac{\frac{R^2_{LMM}}{df_M}}{\frac{1-R^2_{LMM}}{df_R}} \\ \frac{df_M}{df_R} F_{LMM} &= \frac{R^2_{LMM}}{1 - R^2_{LMM}} \\ \frac{df_M}{df_R} F_{LMM} &= \frac{df_M}{df_R} F_{LMM} R^2_{LMM} + R^2_{LMM} \end{aligned}$$

$$R^2_{LMM} = \frac{\frac{df_M}{df_R} F_{LMM}}{1 + \frac{df_M}{df_R} F_{LMM}} \tag{14}$$

In the LMM context, $df_M = p - 1$ for p the number of fixed effects in the model under investigation and df_R could be approximated using various methods, most popularly the Satterthwaite or Kenward-Roger methods[2]. This R^2_{LMM} statistic is then easily derived using the output of standard statistical software for the model of interest.

This statistic measures the ability of the set of fixed effects in the model to explain the variability of the response, which varies in a multivariate space. Practically, it can be used to assess the goodness of a given set of fixed effects immediately without a need to reference other models, though eventually we would like to perform those very comparisons in order to determine which sets of covariates are most valuable to us in a particular context.

Just as we can perform an F-test to measure the adequacy of entire subsets of variables, we can also obtain F-tests measuring the adequacy of individual predictors within each model of interest. That is, instead of testing the multivariate hypotheses in Equation (12) that *all* of the fixed effect parameters are equal to zero simultaneously, we can test if each of the parameters is equal to zero individually. Specifically, we will focus on the Type III F-tests of individual fixed effects parameters. These test for the ability of the individual effect in question to predict the response given that all other effects in the model have been accounted for – i.e., all the other effects are put into the model before the effect is tested.

If we are given a model to work with that has p fixed effects, the hypotheses being tested are,

$$H_{0,j} : \beta_1 | \boldsymbol{\beta}_{(-j)} = 0$$

$$H_{a,j} : \beta_1 | \boldsymbol{\beta}_{(-j)} \neq 0$$

[2] In our application, we use the Kenward-Roger approximation for F-test denominator degrees of freedom where applicable.

Where the notation $\cdot|\beta_{(-j)}$ means *given the all variables other than β_j are already in the model*. The F-test statistic that could be used to test each of the above hypotheses is a special case of Equation (13) in which we set the matrix of contrasts, $\mathbf{C}$, equal to a vector, call it $\mathbf{C}^*$, in which the i^{th} component is 1 and the rest are set equal to zero. Note also that the rank of $\mathbf{C}^*$ is now one, which simplifies the test statistic even further by allowing us to ignore the denominator. If we denote this special case of our F-test statistic by $F_{LMM,partial}$, then we can construct a partial R^2_{LMM} value for each variable in a model given that all the others are already included. That is,

$$R^2_{LMM,partial} = \frac{\frac{df_M}{df_R} F_{LMM,partial}}{1 + \frac{df_M}{df_R} F_{LMM,partial}} \tag{15}$$

This partial R^2_{LMM} measures the amount of multivariate association between the fixed effect in question and the response, given that we have accounted for the other variables in the model. In other words, it looks to address the question: what *additional* proportion of explanatory power can be attributed to this one fixed effect? This is similar to the partial correlation coefficient and partial coefficient of determination in the ordinary multiple regression context [18].

2.3 *t*-Based Linear Mixed Model

The ordinary linear mixed model assumes that both the random effects and the random errors are normal random variables. However, in practice, one might be exposed to data which do not adhere to these assumptions. If one were to perform residual diagnostics and find that there might be more-than-normal variability in a certain type of LMM residual, they might contemplate relaxing the normality assumptions. Here, we consider fitting a model in which we assume that the random effects (e.g., intercepts and slopes) and the errors are distributed as Student's *t* random variables. Pinheiro et al. (2001) [24] propose the following hierarchical model

$$\begin{aligned} \mathbf{Y}_i|\mathbf{b}_i, \tau_i &\overset{indep}{\sim} \mathrm{N}_k\left(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i, \frac{1}{\tau_i}\mathbf{R}_i\right) \\ \mathbf{u}_i|\tau_i &\overset{indep}{\sim} \mathrm{N}_q\left(\mathbf{0}, \frac{1}{\tau_i}\mathbf{G}\right) \end{aligned}$$

$$\tau_i \overset{indep}{\sim} \text{Gamma}\left(\frac{\nu_i}{2}, \frac{\nu_i}{2}\right)$$

Where the new parameters τ_i for $i = 1, \ldots, n$ serve to scale the covariances of both the random effects and the random errors in order to capture any inflated variability due, perhaps, to outlying data points. Under this hierarchy, the LMM can be re-derived to state

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\epsilon}_i$$

Where

$$\mathbf{u}_i \overset{indep}{\sim} t_q(\mathbf{0}, \mathbf{G}, \nu_i)$$
$$\boldsymbol{\epsilon}_i \overset{indep}{\sim} t_k(\mathbf{0}, \mathbf{R}_i, \nu_i)$$

Now, we may model the random effects and errors as being vectors drawn from multivariate Student *t* distributions with degrees of freedom, ν_i, which can be different for any subsets of subject we would like to specify[3]. With this in mind, we turn to the question of how to actually fit this model.

There are two suggested ways to fit this *t*-based LMM: Expectation Conditional Maximization Either (ECME) and Parameter Expanded Expectation Maximization (PXEM). Both are accelerated versions of the canonical Expectation Maximization (EM) algorithm [12].

In ECME, we split the maximization step into several steps, each of which maximizes a subset of parameters while holding all other parameters fixed at their current values. Let $\mathbf{Y}$ denote the complete data, $\mathbf{X}$ denote the observed data, and $\mathbf{Z}$ denote the missing data. Also, let $\boldsymbol{\theta}$ denote our vector of parameters. Take, for example, the case in which we split the parameter set into two separate pieces. Then we would iterate the following two steps.

- E-step: Find $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right) = \mathbb{E}_{\mathbf{Y}}\left(\ell(\boldsymbol{\theta}|\mathbf{Y})|\mathbf{x}, \boldsymbol{\theta}^{(t)}\right)$

[3] In our calculations, we consider the case in which we have a common degrees of freedom parameter for all 66 subjects.

- M-step:
  - Split the parameter vector into a number of bits (say, two of them): $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2]$
  - Find $\boldsymbol{\theta}_1^{(t+1)} = \arg\max_{\theta_1} Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}_1^{(t)}, \boldsymbol{\theta}_2^{(t)}\right)$
  - Next, find $\boldsymbol{\theta}_2^{(t+1)} = \arg\max_{\theta_2} Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t)}\right)$

In the PXEM algorithm Van Dyk (2000) [31] suggests that we consider an expanded parameter space, $\Theta = [\boldsymbol{\theta}, \alpha]$ such that a mapping reduces this expanded space to the parameters in which we are immediately interested, $\mathbb{R}(\Theta = [\boldsymbol{\theta}, \alpha]) = \boldsymbol{\theta}$. This idea is similar in spirit to the idea in ordinary EM that we have a complete data space and a mapping that reduces this space to the observed data, $\mathbb{M}(\mathbf{Y} = [\mathbf{X}, \mathbf{Z}]) = \mathbf{X}$. In the PXEM algorithm, we introduce the augmented data, $\mathbf{Z}$, and parameters, α, over which we may induce whatever convenient properties we like. Then, iteration of the following steps allows us to find the modes of our likelihood (or, as Van Dyk (2000) points out, a posterior in the Bayesian setting).

- E-step: Find $Q\left(\boldsymbol{\theta}, \alpha|\boldsymbol{\theta}^{(t)}, \alpha^{(t)}\right) = Q\left(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)}\right) = \mathbb{E}_{\mathbf{Y}}\left(\ell(\boldsymbol{\Theta}|\mathbf{Y})|\mathbf{x}, \boldsymbol{\Theta}^{(t)}\right)$
- M-step: Find the maximizer of Q w.r.t. $\boldsymbol{\Theta}$: $\boldsymbol{\Theta}^{(t+1)} = \arg\max_{\Theta} Q\left(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)}\right)$
- Use the mapping $\mathbb{R}\left(\boldsymbol{\Theta}^{(t+1)}\right) = \boldsymbol{\theta}^{(t)}$ to eliminate the latent variable α

In our LMM, we introduce the latent variable γ:

$$\frac{\tau_i}{\gamma} \overset{indep}{\sim} \text{Gamma}\left(\frac{\nu_i}{2}, \frac{\nu_i}{2}\right)$$
$$\widehat{\gamma} = \frac{\sum_{i=1}^{n} \nu_i \tau_i}{\sum_{i=1}^{n} \nu_i}$$

so that this new parameter scales the parameters τ_i – which, in turn, scales the original covariance matrices of the random effects and errors. The new parameter uses information concerning the variability of τ_i (in the form of the degrees of freedom parameters ν_i) to, in theory, speed convergence. Once the maximization steps have been derived for the ECME algorithm, the PXEM algorithm uses this new parameter to scale the maximizations in each direction in the parameter space and so is a very natural extension in the context of our LMM problem – e.g., the ECME update for the covariance matrix for the random effects, $\widehat{\mathbf{G}} = \widehat{\mathbf{G}}_{updated}$ now becomes $\widehat{\mathbf{G}} = \frac{1}{\widehat{\gamma}}\widehat{\mathbf{G}}_{updated}$ in PXEM.

2.4 SPECIFICATION OF *a priori* MODELS

In order to properly generate a set of models in which we could invest our attention, we began our analysis by considering all of the valid permutations of our predictors. To begin, ”valid” was defined by setting a few restrictions:

(a) The intercept-only (null) model includes only a fixed intercept, random intercept, and a random effect for distance guessed (i.e., the random coefficients for linear subject-specific response profiles)

(b) All models, except the intercept only model, will include the distance variable as a fixed effect and a random effect

(c) No models can include both the weight and BMI fixed effects simultaneously

(d) No models can include interactions between the image effect and one of weight or BMI effects

Restriction (a) was enforced because we wanted to model each subject as responding linearly across distance guessing occasions. Also, we wished to preserve the covariance structure for every model so that we could directly compare models via an R^2 criterion and because we wished to retain the ability to interpret the subjects’ own response profiles as varying about an overall average profile, as in the effects parameterization in ANOVA. Restriction (b) was made because this specification allowed the population average response profile to vary across distances rather than remaining constant. Restriction (c) was made on the knowledge that, since BMI is a linear transformation of weight, BMI and weight would contain the same information and including them in the same model would only serve to introduce unneeded multicollinearity issues. Finally, restriction (d) was chosen because these interactions were declared *a priori* unimportant practically and would, if included, introduce unnecessary model selection uncertainty.

See Table 2.1 for the full set of models specified.

Table 2.1: All sixteen restricted permutations of the seven fixed effects of interest. M1 is the intercept-only model with a fixed intercept, random intercept, and a random slope term for distance included. *W* denotes the weight variable, *BMI* the body mass index, *I* the body image variable, *D* the distance, and all variables with × are interaction terms.

Model	W	BMI	I	D	$W\times D$	$BMI\times D$	$I\times D$
M1							
M2				■			
M3	■			■			
M4		■		■			
M5			■	■			
M6	■		■	■			
M7		■	■	■			
M8	■			■	■		
M9		■		■		■	
M10			■	■			■
M11	■		■	■	■		
M12	■		■	■			■
M13	■		■	■	■		■
M14		■	■	■		■	
M15		■	■	■			■
M16		■	■	■		■	■

To construct this set of models: we were interested in the distance, weight, BMI, and body image main effects the former three effects' interactions with distance. Enforcing the restrictions (a)-(d) reduced the number of possible models from 1,024 to 130 and imposing the canonical restriction that interactions can only appear in models in which the main effects appear further reduced the possible model set to 16 models. Collaboration confirmed that this set of models contained the models of scientific meaning and interest for the purpose of answering our questions of interest.

3 RESULTS

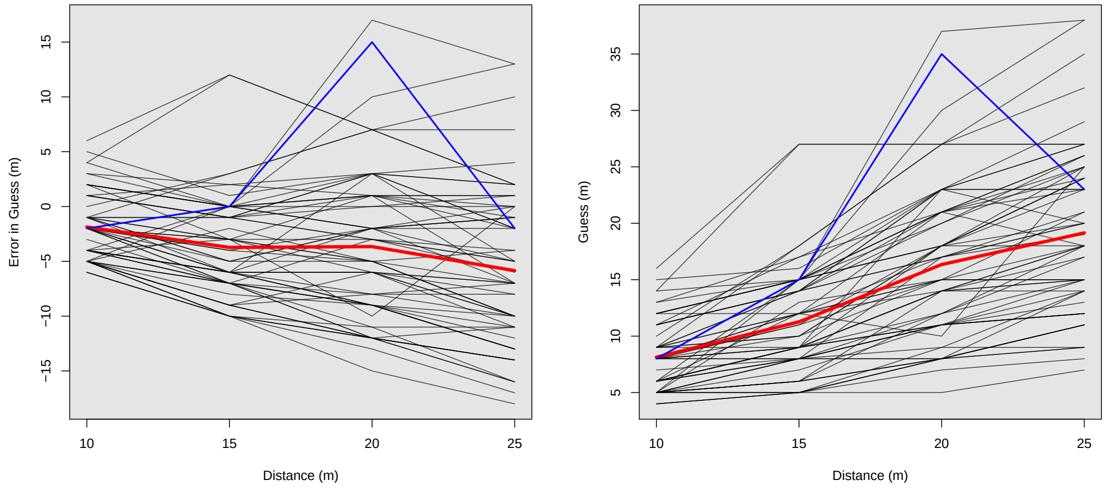
3.1 EXPLORATORY DATA ANALYSIS

To begin our analysis of the distance guessing sample data, we delve into the structure of the sample data via exploratory plots and summaries. First, we include an occasion plot (sometimes called a time plot in repeated measures analyses) in which guesses are plotted against the distance at which the guess was made. In the plot, subjects are represented by one line connecting their guesses at each of the four distances.

The plot is included in Figure 3.1. Panel 3.1a contains a plot of signed error against distance and panel 3.1b plots the actual guesses made by the subjects against distance. These two plots tell us a couple of things straight away. The errors among subjects tend to group near to zero at closer distances and tend to spread out as the cones are moved further away from the subjects. If we take the average error (colored in red in both plots), we would state that the guesses tend to underestimate distances, on average, as cones are moved further away. The last major observation is that there appears to be at least one subject who deviates from the rest by their giving guesses that are far above (about 15 meters above) the actual distances during the 20m and 25m guessing occasions.

This last observation led to a discussion in which it was decided that subject 49 (colored in blue in 3.1) was an anomaly. This subject guessed a very large distance at distance 20m but achieved very little absolute error for all the other distances (possibly a within-subject outlier). It was decided that this subject's data could safely be omitted from the analysis. All other subjects were retained, leaving a total of 66 under investigation.

Next, we observe the relationships between the errors in guesses and each of the predictors individually, depicted in Figure 3.2. Below, ”Total Absolute Error” is equal to the sum of the absolute errors for each subject at every distance to give a one number summary of the overall error that a subject makes. Notice that both weight and BMI appear to be negatively correlated with

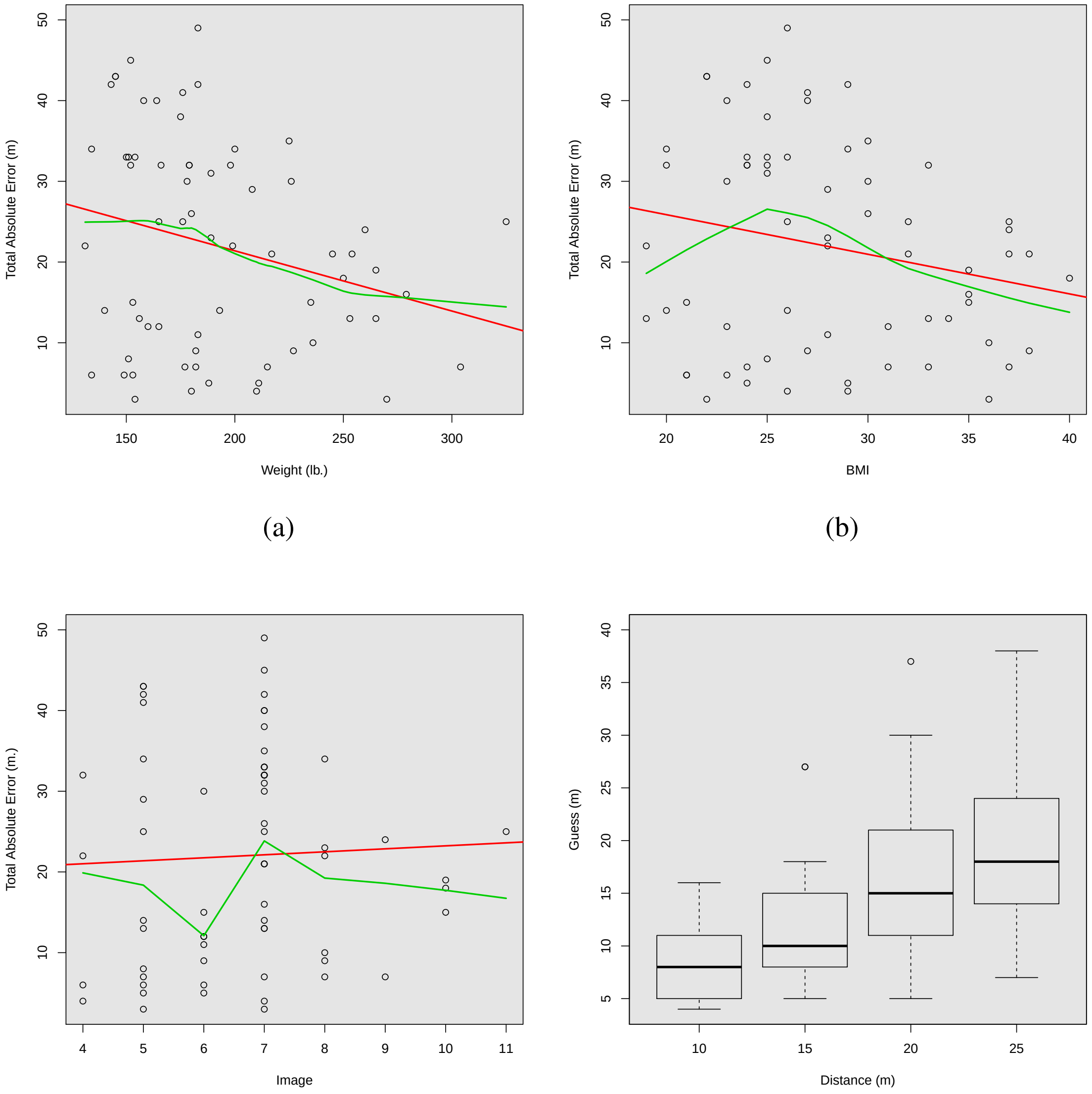


(a) (b)

Figure 3.1: Occasion plot for errors in guessed distances (panel 3.1a) and guessed distances (panel 3.1b) for each subject. Each black line connects responses for a single subject and the red line connects the sample average at each distance. The blue line represents subject 49.

total absolute error. More specifically, it appears that weight is a better predictor of total absolute error by the shape of the LOWESS smooth curve in that it has a steep, consistently downward sloping trend throughout the plot. Contrast this with the LOWESS curve in panel 3.2b in which the curve is positively sloped until BMI is about 26 and decreases from there, though more shallowly than for weight. This indicates to us that BMI weakly explains total absolute error in guesses in comparison to weight. In panel 3.2c, we see that body image seems to be a very weak predictor of total absolute error by the near negligible slope of the simple linear regression line and the overall horizontal trend of the LOWESS smooth curve.

In panel 3.2d, we have side-by-side boxplots of guesses made by all subjects for each distance. The evidence here strengthens our opinions that as distance increases, the guesses decrease relative to the actual distances. We observe that median guesses tend to fall below the actual distances being guessed, prompting us to believe that subjects tend to underestimate with increasingly higher margins of error at further distances. We also observe that variability in guesses increases as we increase distance. We can consult Equation (4) to help realize that we should manipulate the



(a)

(b)

(c)

(d)

Figure 3.2: Plots of weight, BMI, and image against absolute guessing errors in panels 3.2a, 3.2b, and 3.2c, respectively. The red lines are simple linear regression lines and the green lines are LOWESS smoothers on the two variables. The last panel includes boxplots of guesses by distance.

within-subject covariance matrices $\mathbf{R}_1, \ldots, \mathbf{R}_n$, the variance matrix of the random effects $\mathbf{G}$, or both in order to model the heteroskedastic [22] errors across distance guessing occasions.

Lastly, we look at an occasion plot that attempts illustrate the effect of weight and its interaction with distance. We present this in Figure 3.3. Weight was discretized by naming those who fall

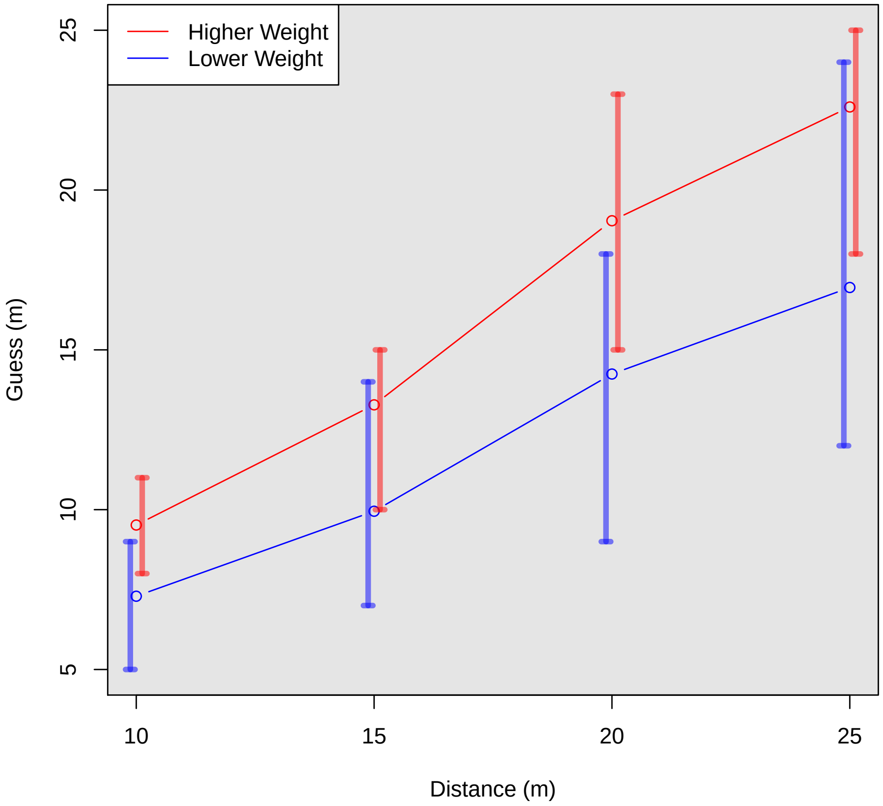


Figure 3.3: Occasion plot for average errors by dichotomized weight over all four distance guessing occasions. The endpoints of the bars at each distance represent the 25% and 75% quantiles of guesses made in the respective weight classes on that distance guessing task.

below the average weight *low* and naming those who fall above the average weight *high*.[4] We then took the average of the guesses made by all subjects falling in each of the weight classes at each of the four distance guessing occasions. Included in the plot are bars denoting the upper and lower quartiles for convenience of comparison. We see that there could be an interaction between weight and distance on guesses made because the average response profiles for the dichotomized weight classes do not change at the same rates across guessing occasions. In fact, it appears as if low weight individuals tend to underestimate the distances more so than do the higher weight

[4] The average weight of the 66 subjects was about 192 lb. The discretization made is done just for convenience of displaying the results and is not carried into the model building stages.

individuals at any distance and the schism between the two groups' average guessing behavior grows as distances get larger.

3.2 Covariance Structure

When constructing a linear mixed model there are two basic components that must be specified: the mean structure (i.e., fixed effects) and the covariance structure (i.e., the random effects and the form of the covariance matrices). One faces a dilemma akin to the chicken and the egg upon the imposition of the choice of which to model first. We choose to model the covariance structure first since our scientific questions of interest focus mainly on the fixed effects in their relations with the response, because an appropriate choice for the covariance structure allows for more precise inference with respect to fixed effects [11], and because model selection focusing on the fixed effects (such as with AICc or R^2_{LMM}) requires a proper covariance structure specification [9].

As for choosing proper random effects before analysis, we have actually already began this iterative process. By our *a priori* model selection, we have decided to employ a linear random coefficients model with random intercepts and random slopes for each subject with respect to distance. What this means in a practical sense is that each subject has some baseline distance guessing behavior unique to themselves (the random intercepts) and we also allow their behavior to change at future guessing occasions (the random slopes).

Our next choice turns us to consider the possible forms of the covariance matrices $\mathbf{G}$ and $\mathbf{R}_1, \ldots, \mathbf{R}_n$. To go about our decision-making, we will make use of two approaches. The first will involve an information-based procedure in which we will use AICc to choose among proper forms the $\mathbf{G}$ matrix, then among proper forms of the $\mathbf{R}_i$ matrices, and finally among the two simultaneously. To bolster any final decisions, we will rely on likelihood ratio tests, with an eye towards caution before testing non-nested covariance structures.

By our discussion among the estimation methods, ML and REML, in Section 1.4 we choose to use the REML estimation method and hold all of the fixed effects constant while we are testing different forms of the covariance matrices. This begets our choice of fixed effects to include while we are making these comparisons. Our comparisons of covariance structures are made with the best AICc model, M11, from Table 2.1. In addition, we impose a variance components (VC) form – that is, $\mathbf{R}_i = \sigma_\epsilon^2 \mathbf{I}$ – on the within-subject covariance matrices for these initial derivations.

First, we consider the $\mathbf{G}$ matrix, or the covariance matrix of the random coefficients. Using SAS 9.4, we recover the information as presented in Table 3.1. We point the reader to the SAS 9.2 manual [26], Fitzmaurice et al. (2011) [11], or another book with a dedicated linear mixed model section for a full review of the various covariance matrix forms presented in the table. The default form of variance components was ranked fourth in terms of AICc and so we should specify a better form in order to reap the benefits of precision gains in inference. The top two competing forms by AICc alone are the unstructured – allowing a different covariance parameter for every element of the matrix – and the heterogeneous compound symmetry – specifying one parameter for each diagonal element.

Table 3.1: Variance forms imposed on the covariance matrix of the random coefficients using the fixed effects from model 11 in Table 2.1. Parameters denotes the number of covariance parameters that must be estimated in the model. -2ℓ refers to twice the negative log likelihood derived from the model. Italicized entries denote model fits that result in non-positive-definite covariance matrices.

Form of $\mathbf{G}$	AICc	Parameters	-2ℓ
Heterogenous Compound Symmetry (CSH)	*1,366*	*3*	*1,360*
Unstructured (UN)	*1,369*	*4*	*1,360*
Toeplitz (TOE)	*1,393*	*3*	*1,387*
Variance Components (VC)	1,393	3	1,387
Compound Symmetry (CS)	*1,431*	*2*	*1,427*
Lag-1 Autoregressive (AR(1))	*1,431*	*2*	*1,427*

If were to naïvely perform a likelihood ratio test among forms of the $\mathbf{G}$ matrix, we could appeal to the asymptotic chi-squared distribution of the deviance among the two models. However, this would overlook a glaring problem: the model fits that are italicized in Table 3.1 result in covariance matrices that are not positive-definite. That is, all but the variance component form of $\mathbf{G}$ resulted in model fits that were unstable and did not give reliable results within which we could place our confidence. Even after a series of workarounds, the problems persisted. These complications could be due to the moderate sample sizes (66 subjects to estimate between-subject effects and 4 observations per subject to estimate within-subject effects). We therefore ended up declaring the only reasonable choice to be the VC form and proceeded to check for alternative forms of the within-subjects covariance matrices.

We again use the fixed effects from M11 from Table 2.1 and use REML estimation methods, and we now set $\mathbf{G}$ as a VC matrix. The results are listed in Table 3.2 and the reader will notice that we have omitted the CS and TOE forms, which is due to non-convergence of the estimation procedures. With these two eliminated, our top two competitors in terms of AICc are again CSH and UN. Also, again, we see that both these forms resulted in fits that were not stable and resulted in non-positive-definite covariance matrices. Therefore, we opted to choose a VC form for the within-subject matrices as well due to the apparent instability.

Table 3.2: Variance forms imposed on the within-subject covariance matrices using the fixed effects from model 11 in Table 2.1. Parameters denotes the number of covariance parameters that must be estimated in the model and -2ℓ refers to twice the negative log likelihood derived from the model. Italicized entries denote model fits that result in non-positive-definite covariance matrices.

Form of $\mathbf{R}_i$	AICc	Parameters	-2ℓ
UN	*1,365*	*12*	*1,340*
CSH	*1,368*	*6*	*1,356*
VC	1,393	3	1,387
AR(1)	1,393	6	1,385

In the end, we have decided upon modeling $\mathbf{G}$ as a variance components matrix,

$$\mathbf{G} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

And the within-subject matrices as variance component matrices as well,

$$\mathbf{R}_i = \sigma_\epsilon^2 \mathbf{I} = \begin{bmatrix} \sigma_\epsilon^2 & 0 & 0 & 0 \\ 0 & \sigma_\epsilon^2 & 0 & 0 \\ 0 & 0 & \sigma_\epsilon^2 & 0 \\ 0 & 0 & 0 & \sigma_\epsilon^2 \end{bmatrix}$$

In all, these specifications require 3 variance parameters to be estimated, which we believe is a reasonable choice given the moderate size of the data set. In our search for proper forms of both types of covariance matrices simultaneously, we arrive at the same conclusion: VC structures are the most appropriate for both given our sample data.

3.3 MEAN STRUCTURE

We now focus on the mean structure of the linear mixed model that we plan to fit to our sample data. Recall that our inference will be based on model averaged estimates and so we do not need to choose a single mean structure, but are only required to rank them in terms of relative information content. Chen, et al. (2013) [8] beautifully stated that model averaging ”can be thought as a continuous extension of model selection” in that we will not be assigning discrete, binary values to models based on their worth. Instead, we will assign values that may range continuously between zero and one – restricted to sum to one – to each model so that multiple data sources may be used in inference.

To begin, we rank the sixteen models chosen *a priori* from Table 2.1 according to their AICc taking the methods from Section 2.1. The results for our sixteen chosen models are included in Table 3.3. All of our models were estimated using ML estimation in SAS 9.4 (see the discussion in Section 1.4).

The results in Table 3.3 illuminate much about the problem at hand. The first thing that we notice is that M11 has the lowest AICc value and includes the fixed intercept, the main effects for

Table 3.3: Values used to compute AICc model weights w. Unique model numbers are included on the left, followed by the terms included in the model. Raw AICc values (lower is better) lie to the right of the model terms. The ΔAICc terms are the difference in AICc between each model's measurement and the lowest AICc in the *a priori* set.

Model	Model Terms	AICc	ΔAICc	w
M11	(Intercept)+W+I+D+W×D	1,380	0.00	0.43
M8	(Intercept)+W+D+W×D	1,381	0.75	0.29
M15	(Intercept)+W+I+D+W×D+I×D	1,382	1.98	0.16
M6	(Intercept)+W+I+D	1,385	4.93	0.04
M13	(Intercept)+W+I+D+I×D	1,385	5.04	0.03
M3	(Intercept)+W+D	1,386	5.73	0.02
M9	(Intercept)+BMI+D+BMI×D	1,388	8.16	0.01
M12	(Intercept)+BMI+I+D+BMI×D	1,388	8.56	0.01
M4	(Intercept)+BMI+D	1,389	9.29	0.00
M7	(Intercept)+BMI+I+D	1,389	9.65	0.00
M14	(Intercept)+BMI+I+D+I×D	1,390	9.79	0.00
M2	(Intercept)+D	1,390	10.33	0.00
M16	(Intercept)+BMI+I+D+BMI×D+I×D	1,390	10.69	0.00
M5	(Intercept)+I+D	1,392	12.22	0.00
M10	(Intercept)+I+D+I×D	1,392	12.34	0.00
M1	(Intercept)	1,490	110.20	0.00

distance, weight, and image, and an interaction effect between distance and weight. With probability 0.43, this is the best model in the *a priori* set using our sample data in terms of information retained. The second best model, M8, has all of the same effects as M11 except for the main effect for image and has an Akaike weight of 0.29. Collectively, the top five models by AICc have weight about 0.95, indicating that the remaining eleven models will have relatively little say in terms of the model averaged point estimates – though their inclusion is crucial in accounting for model selection uncertainty. Also, all five top models include weight as a predictor while the first model with BMI as a predictor is ranked number seven in terms of AICc, which demonstrates substantial evidence for weight being a more appropriate predictor of guessed distance.

In addition to ranking the models by corrected Akaike weights, we also chose to rank our models by the generalized R^2 criterion for linear mixed models reported in Equation (14). Recall that the R^2_{LMM} values are dependent upon the overall F-test for the set of fixed effects included in the models and also upon the denominator degrees of freedom used in constructing the F-test statistics. In this situation, we have chosen to use the Kenward-Roger estimates for the denominator degrees of freedom. The results are included in Table 3.4[5].

Table 3.4: R^2_{LMM} terms for the explanatory ability of sets of fixed effects in the fifteen non-null *a priori* models. Also included are the model terms and each model's rank by AICc for comparison with Table 3.3.

Model	Model Terms	R^2_{LMM}	Rank by AICc
M2	(Intercept)+D	0.79	12
M8	(Intercept)+W+D+W×D	0.76	2
M9	(Intercept)+BMI+D+BMI×D	0.74	7
M11	(Intercept)+W+I+D+W×D	0.74	1
M3	(Intercept)+W+D	0.74	6
M6	(Intercept)+W+I+D	0.73	4
M15	(Intercept)+W+I+D+W×D+I×D	0.73	3
M10	(Intercept)+I+D+I×D	0.73	15
M4	(Intercept)+BMI+D	0.73	9
M7	(Intercept)+BMI+I+D	0.73	10
M5	(Intercept)+I+D	0.73	14
M13	(Intercept)+W+I+D+I×D	0.72	5
M12	(Intercept)+BMI+I+D+BMI×D	0.72	8
M16	(Intercept)+BMI+I+D+BMI×D+I×D	0.71	13
M14	(Intercept)+BMI+I+D+I×D	0.71	6

We see that the model with the main effect for distance as the only covariate, M2, has the highest R^2_{LMM} of about 0.79. This means that distance explains about 79% of the multivariate variability in the response. Any additions of covariates only serve to decrease the overall explanatory power of the fixed effects. For example, the best AICc model, M11, is ranked fourth with respect to this measure of goodness.

Since both the AICc and R^2_{LMM} values are to be used to compare among models, we include a column of ranks by AICc in Table 3.4. When we consider the patterns in the R^2_{LMM} values, we

[5] Note that the fixed intercept-only model, M1, is not included because the null model is that which has no fixed effects of immediate interest to us and serves as the reference model against which all others are tested.

observe that, barring M2, four of the top five models include weight as a predictor and one includes BMI. This bolsters our observation from before that weight tends to explain more of the variability in the response over BMI. In addition, it appears that body image is scattered about the models and so gives little to no indication of its worth as a predictor of distance guessing ability.

To take our analysis of the R^2_{LMM} statistic even further, we turn our attention to the partial measures of multivariate explanatory power from Equation (15). The complicating factor in investigating these partial measures is that we can calculate one $R^2_{LMM,partial}$ for every variable from every model. This leaves, for example, fifteen values for the distance variable alone. We simplify these measures by taking the simple mean of the $R^2_{LMM,partial}$ across all sixteen models for each of the non-intercept fixed effects and include the results in Table 3.5.

It is important to note that taking averages of these statistics across models complicates their interpretations since they are no longer estimates of the explanatory power of the variables given the others included in the model. Since we have averaged over several models, and every model contains a different set of fixed effects, the ”given the others” interpretation varies from model to model. In an attempt to find common ground between rigor and parsimony, we urge the reader to think of these values as rough measures of the explanatory ability attributable to individual fixed effects, accounting for the other variables declared important *a priori*.

Taking a look at the results, we see that the distance variable has an averaged partial coefficient of determination of $\bar{R}^2_{LMM;partial} \approx 0.34$. We will take this to mean that about 34% of the multivariate variability in the response tends to be explained by the distance variable once we take all the other fixed effects into account (since distance appears in models including all six other fixed effects). Also, weight tends to explain about 4% of the variability in the response given all the other variables except BMI (since weight appears in models with all other variables save for BMI). Measures such as these will be instrumental in interpreting the worth of specific fixed effects.

Table 3.5: Simple averages of partial R^2_{LMM} values for each of the non-intercept fixed effects taken with respect to all other variables in each of the respective models.

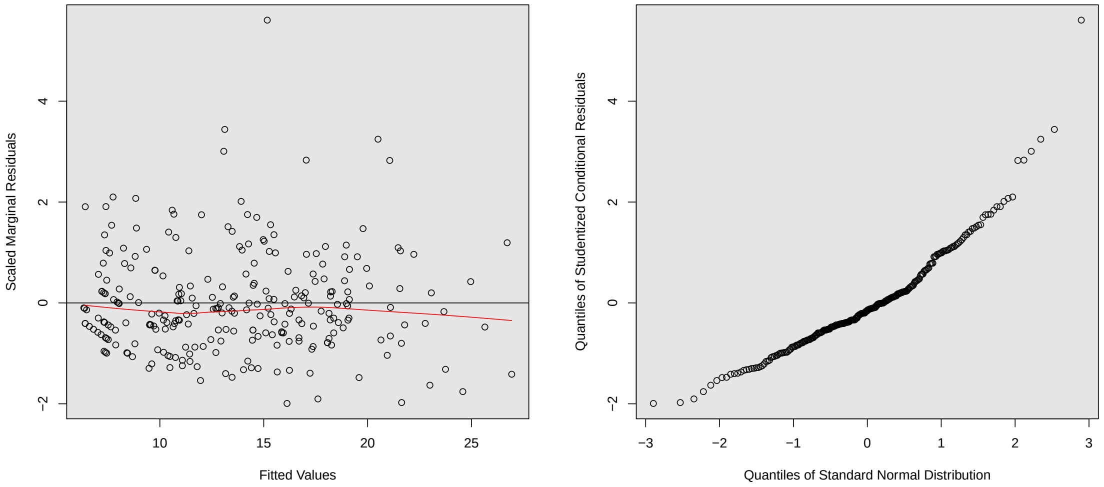
Effect	Partial $\bar{R}^2_{LMM}$
Distance	0.34
Weight	0.04
BMI	0.02
Image	0.02
Weight × Distance	0.02
BMI × Distance	0.01
Image × Distance	0.01

3.4 Model Diagnostics

We now turn to check the model assumptions via the diagnostic tools discussed in Section 1.5. The complicating factor in our model diagnostics is that we do not eliminate all but one model with which we plan to perform inference. Instead, we retain all sixteen models decided to be *a priori* important. Thus, we have sixteen models on which we should perform diagnostics. This has been done, but for our purposes in this report, we restrict ourselves to discussing any diagnostic results with respect to one model and that model is the best-by-AICc model, M11.

We choose to do this because, if we were performing an ordinary analysis rather than adhering to the model averaging framework, then this would be the model on which we would be performing diagnostics. In addition, M11 is one of the more saturated models and we believe that it is representative of the forms of models we are considering. In fact, after checking all of the model diagnostic plots, we notice that all sixteen models' diagnostic plots take similar forms and are almost indistinguishable amongst themselves, which adds credence to both the model building process and to the diagnostics to come.

To begin, we consider the marginal residuals, which can be used to check for the adequacy of the mean structure induced by the fixed effects and for analyzing our choice of within-subject covariance matrices. We use the scaled marginal residuals in the two plots included in Figure 3.4. In panel 3.4a, we see that there is no trend apparent in the residuals, which supports the hypothesis



(a)

(b)

Figure 3.4: Residual diagnostic plots using the scaled marginal residuals from M11. Panel 3.4a contains the scaled residuals plotted against the fitted values. The black line is the horizontal line at zero and the red line is a LOWESS smooth line of the residuals, both for visual reference. Panel 3.4b is a normal QQ plot of the residuals.

that our choice of fixed effects is adequate in describing the sample data. There is little evidence of heteroskedasticity and so our model seems to account for within-subject variability well enough. One observation, with a scaled residual above 4.5, appears to deviate from the rest and could be considered an outlying value with respect to the overall mean response profile (we will return to this below). In panel 3.4b, we have a normal QQ plot of the scaled marginal residuals. Other than some deviant behavior about the upper tail, promoted by the previously mentioned observation, nothing seems to strongly indicate any lack of normality of the marginal residuals.

Similar plots of the studentized conditional residuals are included in Figure 3.5. Recall that the conditional residuals can be useful in diagnosing issues in the choice of the overall mean structure as well as the choice for the structure of the response covariance matrix, $\mathbf{V}$. In panel 3.5a, we do not notice any obvious trend and this is supported by the slope of zero in the LOWESS smooth curve. However, there seems to be increasing error variance as the fitted values increase and there appear to be potential outliers for higher fitted values. Panel 3.5b shows that the distribution of the conditional residuals have heavier tails than the normal distribution, which leads us to question this assumption.

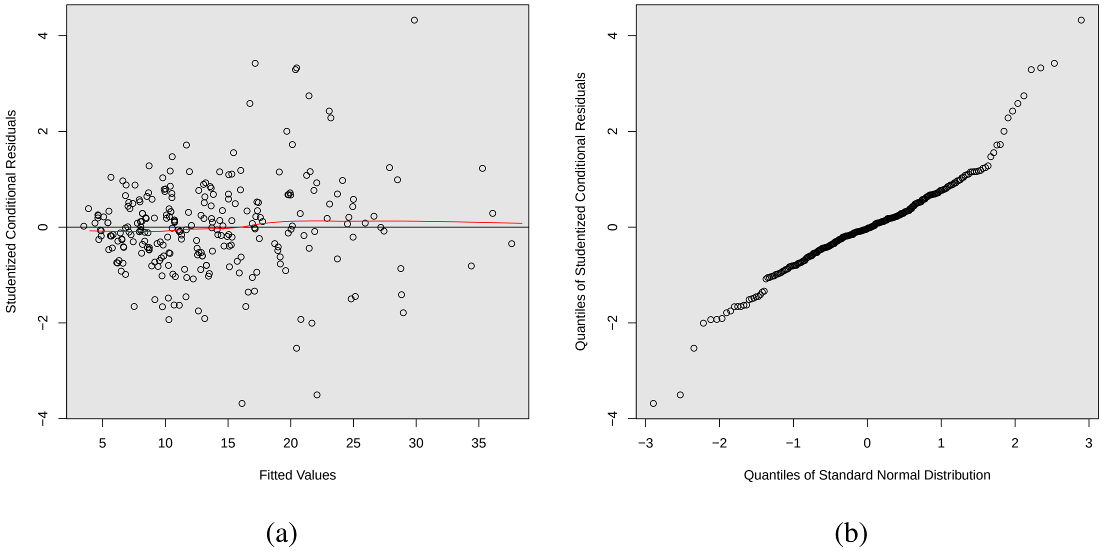


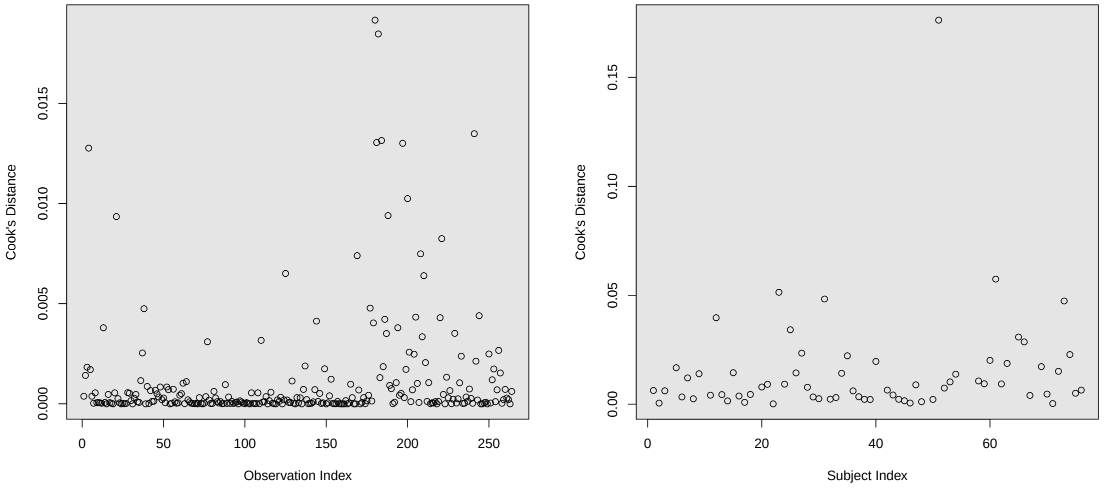
Figure 3.5: Residual diagnostic plots using the studentized conditional residuals from M11. The plotting aspects of these plots are the same as in 3.4.

Influence diagnostics are performed using Cook's distances for both observations and for subjects in Figure 3.6. As expected, we have a couple of individual observations that have Cook's distances that are relatively far above the rest (the pair of points at Cook's distances around 0.015). These observations are to be noted as potentially influential points. But, if we are to adhere to the general rule that values above either of the 25^{th} or 50^{th} F quantiles then we would not consider any of the observations to be influential to the estimation of the fixed effects and could attribute their large Cook's distances to sampling variability..

In panel 3.6b we have influence measures for entire response profiles of subjects. We see the same sort of pattern here in which we have one potentially influential subject, though when we consider the F percentiles then none of the subjects appear to significantly influence our estimation of the fixed effects[6].

Taking the observations from Figures 3.4, 3.5, and 3.6, we would be confident in saying that our choice of mean structure is adequate. However, we see some problems in the apparent heteroskedasticity among the conditional residuals. That is, after accounting for the fixed effects and

[6] F percentiles for outlying observations are $F_{(0.25;5,264-5)} \approx 0.53$ and $F_{(0.50;5,264-5)} \approx 0.87$ and are $F_{(0.25;5,66-5)} \approx 0.53$ and $F_{(0.50;5,66-5)} \approx 0.88$ for outlying subjects.



(a) (b)

Figure 3.6: Cook's distances for individual observations, 3.6a, and for subjects themselves, 3.6b.

the random slopes and intercepts, we still have guessing errors that vary more about their respective means at increased distances. In addition, the distributions of the conditional residuals don't seem to have strong evidence of being normal. The presence of the potentially influential observations also spurs us to consider the fact that our sample data might not support our distributional assumptions.

3.5 Remedial Measures

3.5.1 Power of X

In order to reconcile the possible heteroskedasticity seen in Figure 3.5, we implement a power of X model. This model is recommended by Littell et al. (2006) [20] as a way to account for variability of errors being functions of one or more covariates. In effect, this model makes a single adjustment to the error variance.

Instead of defining the random errors $\mathrm{Var}(\boldsymbol{\epsilon}_i) = \mathbf{R}_i = \sigma_\epsilon^2\mathbf{I}$ for $i = 1, \ldots, k$, we instead define $\mathrm{Var}(\boldsymbol{\epsilon}_i) = \mathbf{R}_i = \sigma_\epsilon^2 \exp\{\gamma \mathbf{x}_{ji}\}\,\mathbf{I}$. Here, the exponential function acts component-wise along the vector of the j^{th} covariate for the i^{th} subject, $\mathbf{x}_{ji}$. This operation results in within-subject covariance matrices that explicitly allow for scaling of the variability of the response when moving along the direction of a covariate of the user's choice. In addition, this scaling is additionally controlled by the parameter γ.

In application of this new model to our sample data using M11, we naturally chose the "X" in the power of X to be distance. This is because such a choice was suggested by the diagnostic plots given above and because it is the only variable to vary within subject in this example. The estimated dispersion parameter $\widehat{\gamma} \approx 0.03$ with variance about 0.03. A Wald-based normal p-value for the test of this parameter equalling zero was about 0.2. However, since this is a test on the boundary of the parameter space then the regularity conditions required by normal-based likelihood tests are violated and so the Wald test is not valid in this case.

Instead, we can use an augmented likelihood ratio test for the hypotheses $H_0 : \gamma = 0$ versus $H_a : \gamma \neq 0$. Some authors approach the problem as follows. While attempting to derive the sampling distribution of $\widehat{\gamma}$ under the null hypothesis, we can think of two cases. By natural variability in the data, we can think of $\widehat{\gamma}$ being estimated at some non-zero number by random chance alone. In this case, there is one additional parameter to be estimated in the power of X model compared to the model without the power of X covariance structure. On the other hand, one may think of a situation where $\widehat{\gamma}$ is really being estimated at exactly zero. In this second case, there would be no additional parameters being estimated under the model. Then, if we think of these two cases as being equally probable, the sampling distribution of $\widehat{\gamma}$ can be thought of as taking on the form of a mixture of a χ^2 distribution with one degree of freedom half of the time and a χ^2 distribution with zero degrees of freedom the other half of the time [11]. Therefore, our likelihood ratio test statistic should then be compared to the quantiles of this mixture distribution instead of the $\chi^2_{(1)}$ distribution.

Since the difference in twice the negative log likelihoods for the ordinary M11 and for the power of X M11 was,

$$1,363.2 - 1,359.6 \approx 3.6$$

and since the 0.025 and 0.05 quantiles of the 0-1 mixture of χ^2 distributions are about 3.84 and 2.71 then our likelihood ratio test $0.025 \leqslant$ p-value $\leqslant 0.05$. Therefore, at the 0.05 significance level, one would reject the null hypothesis and would prefer the power of X version of M11 over the ordinary version. That is, the added term for adjusting the covariance based on measurement occasion is significantly different from zero (if only marginally). This significance is borderline and because residual plots for the power of X model show very much the same heterskedastic patterns as seen in Figure 3.5, we would not have gained much at all by adopting this new model. Because the power of X model would only serve to complicate the model while not mending the issue we set out to address in the first place, we continue to utilize the models without the exponential structure in our analysis in order to preserve parsimony and interpretability in this context.

3.5.2 *t*-Based Linear Mixed Model

For our distance-guessing data, we attempted to fit M11 in three ways: (1) the ordinary, normal-based LMM using the MIXED procedure in SAS 9.4, (2) the *t*-based LMM using ECME, and (3) the *t*-based LMM using PXEM. For the *t*-based model fits, we were forced to code everything by hand, but were able to use derivations from the Pinheiro, Liu, and Wu (2001) paper to aid in our coding.

Table 3.6: Fixed effects estimates using the estimation procedures employed by SAS under the normal LMM and using ECME and PXEM under the t-based LMM. The PE values are point estimates for the fixed effects and the StDev values are their estimated standard deviations.

	SAS		ECME		PXEM	
Effect	PE	StDev	PE	StDev	PE	StDev
(Intercept)	5.56	1.99	4.16	1.58	4.16	1.58
Distance	0.21	0.20	0.22	0.16	0.22	0.16
Weight	0.03	0.01	0.03	0.01	0.03	0.01
Image	−0.60	0.35	−0.46	0.26	−0.46	0.26
Weight×Distance	0.003	0.001	0.002	0.001	0.002	0.001

Table 3.7: Covariance parameter estimates using the estimation procedures employed by SAS under the normal LMM and using ECME and PXEM under the t-based LMM.

Effect	SAS	EMCE	PXEM
G Matrix	$\begin{bmatrix} 8.00 & 0 \\ 0 & 0.11 \end{bmatrix}$	$\begin{bmatrix} 3.87 & 0.41 \\ 0.41 & 0.05 \end{bmatrix}$	$\begin{bmatrix} 3.87 & 0.41 \\ 0.41 & 0.05 \end{bmatrix}$
$\mathbf{R}_i$ Matrices	$3.94 \cdot \mathbf{I}_4$	$3.96 \cdot \mathbf{I}_4$	$3.96 \cdot \mathbf{I}_4$

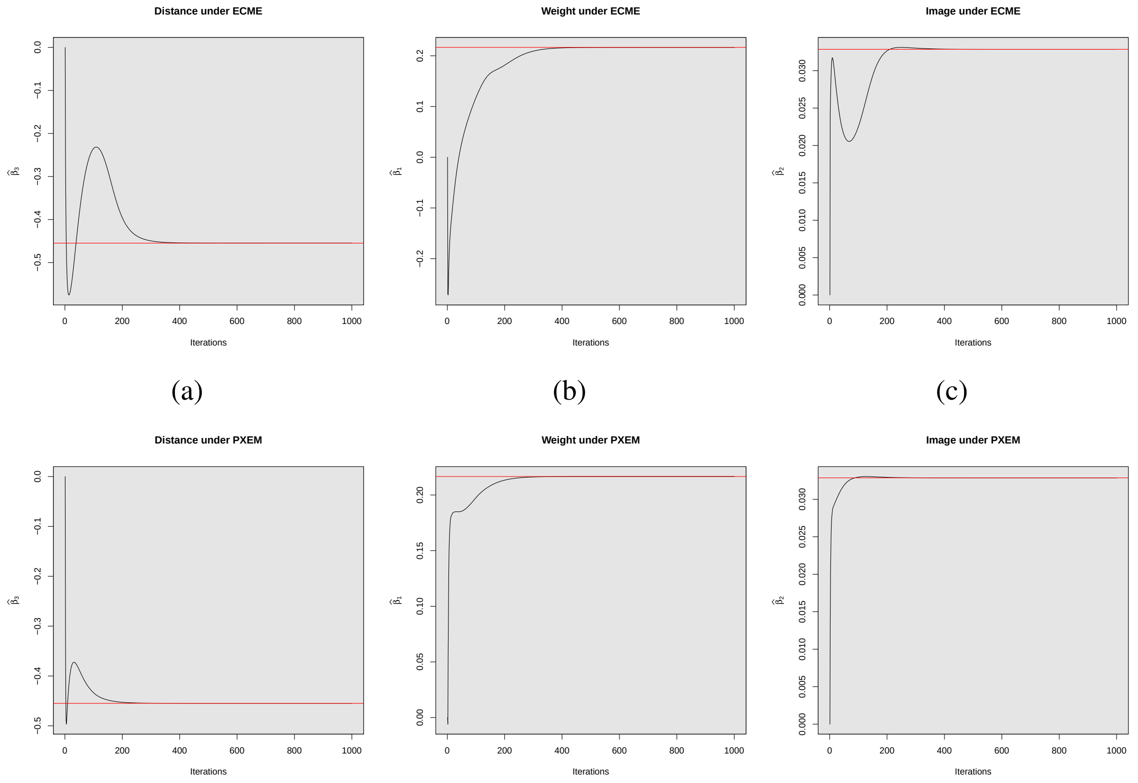
Results for the fixed-effects parameter estimates can be found in Table 3.6. The largest differences between the two model fits are the estimates for the intercept and the body image effect. The other point estimates are almost identical to one another across the normal- and t-based models. It appears that under the *t*-based model the standard errors have been reduced, indicating that using the new model allows us to have slightly more precise estimation. Take, for example, the distance effect. Under the normal LMM, the point estimate was about 0.21 with standard error about 0.20 while under the *t* LMM, the estimate has changed to about 0.22 with decreased standard error 0.16. The ECME and PXME fits agree on all of the estimates.

Comparisons of the estimates for the covariance parameters using the various model fitting procedures can be found in Table 3.7. It appears as if the error variances are comparable among the ordinary and rubust model fits. However, we find some notable differences in the between-subject matrices. First of all, the variability of the subject-specific baseline guesses decreases from 8.00 to about 3.87. Also, the variability of the subject-specific slopes changes from about 0.11 to about 0.05. These two variances are essentially halved, so why might this be? Reason number one includes the non-zero off-diagonal entry. In the SAS fit, we restrict the covariance between the random intercepts and slopes to be zero due to unstable fits in SAS. However, the data-driven robust fitting procedures allow for non-zero covariances among the random coefficients and this small change allows for changes to the variance estimates as well. The second reason harks back to the decreased standard deviations we saw when considering the fixed effects. We tend to see these decreases because the t-based model seems to fit the sample data better, which leads to increased precision.

In addition to the information in Tables 3.6 and 3.7, the hierarchical model can be used to diagnose the need for the t distributional assumptions through the estimate of the degrees of freedom, ν. This degrees of freedom parameter is estimated in the EM algorithms using the data to tailor the form of the t distributions. If the normal model were appropriate, we would estimate that $\nu >\sim 30$. However, the data-based estimates for the degrees of freedom were about 3.20 under ECME and 4.07 under PXEM, indicating a possible need for the increased variability of between-subject and within-subject deviations allowed by the t distributions.

Figure 3.7 illustrates the speed of convergence of three of the fixed effects' point estimates using the ECME and PXME algorithms ran with 1,000 iterations. From these plots, we see that the PXEM algorithm does, indeed, have a speed advantage over the other. The ECME method converged to its end estimates at around 300 iterations for all of the estimates while the PXEM method converged around 200 iterations. While this did not present a whole lot of practical difference with respect to computing time in our relatively simple model (about 5 minutes on a 2 GHz dual-core MacBook running OSX v10.6.8 on 3 GB of RAM for 1,000 iterations), we could see the potential advantages in saving 100 or so iterations of computing time elsewhere.

Based on the output in Table 3.6, we do not foresee the outcomes of our analysis being any different if we use the *t*-based models over the standard, normal-based models. Thus, even though it is suggested that we might want to use the *t*-based model (recall the small estimated degrees of freedom and the reductions in uncertainty estimates) to account for non-normal variability, we revert to using the normal-based LMM in the rest of this report since we believe that any conclusions we make will be the same no matter which model we use, because our inference is restricted to explanation and are more concerned with parsimony and clear demonstration of the important covariates – so we do not necessarily desire every reduction in uncertainty estimates as we might want in an application focused on prediction.



(a) (b) (c)

(d) (e) (f)

Figure 3.7: Convergence plots for the parameter estimates for the weight, distance, and image main effects in the *t*-based M11 using ECME in panels 3.7a-3.7c and PXEM in panels 3.7d-3.7f. The red lines depict the final values taken for each of the parameter estimates.

3.6 Model Averaged Estimates and Inference

Having decided to use the normal-based LMM using ML estimation for fixed effects *and* using ML estimation for covariance parameters, we fit all 16 models from Table 2.1. Then, to generate model averaged estimates we used Equations (10) and (11) along with the model weights computed in Table 3.3 to calculate the point estimates and standard errors for the fixed effects in Table 3.8.

The bolded entries in the table indicate fixed effects that are declared significantly different from zero at the 95% confidence level using Wald intervals. These intervals use the variances unconditional on the model being used and thus account for the uncertainty imbedded in the model selection process.

Surprisingly, distance is not a significant effect with 95% confidence. As distance is increased, we tend to see increases in guessed distances with a mean of 0.27 meters for every meter increase in actual distance – with 95% confidence between about -0.20 and 0.73 meters – when we change only the distance being guessed. We would have expected distance to be a significant effect, since as the actual distance increases we would expect guessed distances to increase as well. However, as more covariates are included in the model, the effect of increasing cone distance gets overshadowed by the effect of weight and others during model averaging.

Table 3.8: Model averaged estimates, their unconditional standard deviations, and Wald 95% confidence intervals for all eight fixed effects. Bolded entries are significantly different from zero with 95% confidence.

Effect	PE	StDev	Wald 95% CI
Intercept	**5.04**	**2.06**	**(0.99, 0.91)**
Distance	0.27	0.24	(-0.20, 0.73)
Weight	**0.028**	**0.014**	**(0.002, 0.06)**
BMI	0.0047	0.0098	(-0.0145, 0.0239)
Image	-0.40	0.41	(-1.19, 0.40)
Weight × Distance	0.0025	0.0013	(-0.0001, 0.0051)
BMI × Distance	0.0002	0.0005	(-0.0007, 0.0012)
Image × Distance	0.0023	0.0042	(-0.0060, 0.0105)

Weight is the only non-intercept fixed effect that is significantly different from zero at our chosen confidence level. One interpretation of this could be that if two subjects differ only in weight by 100 pounds, we expect to see differences in their guessed distances by about 2.8 meters – between about 0.2 and 6 meters with 95% confidence. Specifically, the heavier subject will tend to guess the longer distance and the lighter subject the shorter. The confidence interval around this estimate is relatively wide, ranging about 6 meters. Nevertheless, the effect of heavier individuals reporting longer distance guesses regardless of distance guessing occasion persists in this sample.

The weight by distance interaction would be significant at the 90% confidence level, but certainly is *not* at the 95% confidence level. The meaning of this effect is that, for subjects differing in weight by 100 pounds only, we tend to see differences in guesses of 2.8 meters at the baseline distance and expect to see them deviate from one another by about 1.3 meters at each distance guessing occasion – that is, an increase in cone distance of 5 meters. Although not significant at the 0.05 significance level, we can observe the evanescent interaction effect visually.

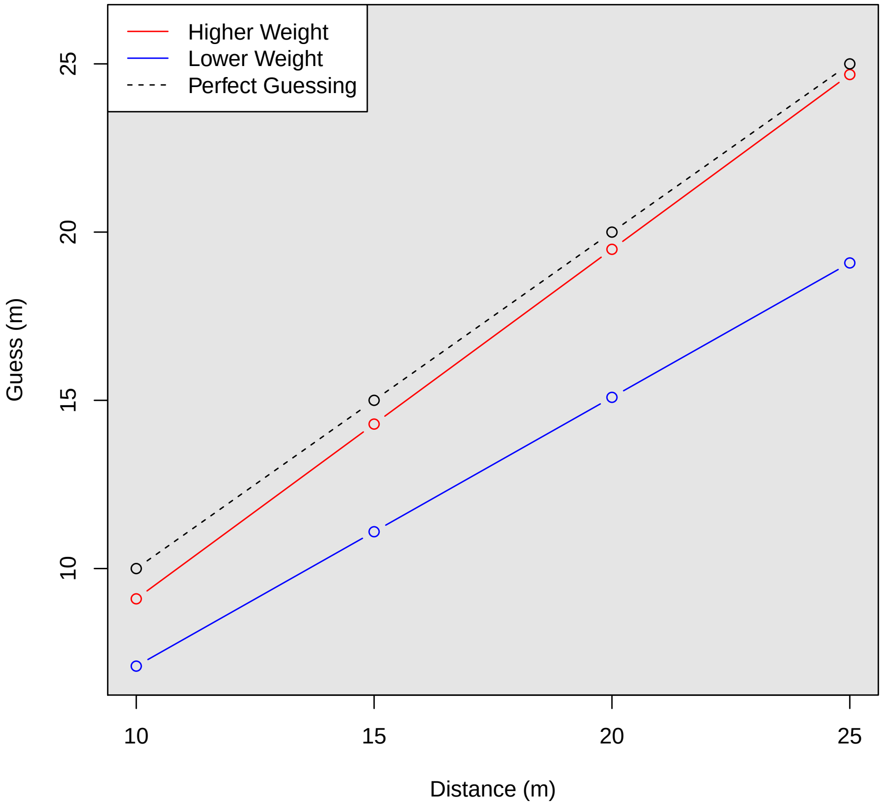


Figure 3.8: Illustration of the model averaged predicted response profiles for subjects differing only in their weights, in which they differ by 100 pounds. The lower-weighted individual is depicted in blue and the higher-weighted individual in red. The black dashed line depicts an ideal subject with perfect guesses at each distance.

Figure 3.8 depicts the model predictions from the model averaged estimates for average response profiles for subjects that differ by 100 pounds – the subjects are the same in all other respects including BMI and their body image and is reflective of comparing subjects 1 and 2 from the actual study. Notice that as the actual distance is increased, we tend to see the two response profiles deviate from one another, just like we saw in the sample data in Figure 3.3. This effect, and main effect of weight itself, results in higher-weighted individuals making more accurate guesses at all of the distance guessing occasions.

Using similar model averaging techniques, we include the model averaged point estimates, their unconditional standard deviations, and Wald-based 95% confidence intervals for the covariance parameters in Table 3.9. All of the effects are significant at the 0.05 significance level. We do note that the Wald-based confidence intervals presented for the covariance parameters should be taken with a grain of salt due to the fact that proper tests of significance for covariance parameters being different from zero should be based on likelihood methods and, in particular, the mixture of χ^2 distributions should be used in these tests as done in Section 3.5.1. In either case, we derive the following inference from the results in Table 3.9.

Table 3.9: Model averaged estimates, their unconditional standard deviations, and Wald 95% confidence intervals for random coefficients covariance parameter estimates.

Effect	PE	StDev	Wald 95% CI
$\widehat{\sigma}_1^2$	8.15	1.90	(4.43, 11.87)
$\widehat{\sigma}_2^2$	0.11	0.02	(0.06, 0.15)
$\widehat{\sigma}_\epsilon^2$	3.94	0.45	(3.07, 4.81)

The first thing of note in the table is the sheer magnitude of the variability of the subject-specific intercepts. In other words, the subjects come into the study with vastly different guessing behaviors at the baseline guessing occasion. Specifically, participants differ in their guesses at baseline by about 3 meters on average, which rivals the distance the cone is moved at the next guessing occasion! This fact remains apparent if we consult the plot of raw response profiles in Figure 3.1 in which we see baseline guesses at the 10 meter cone ranging from about 6 meters to about 16.

We tend to see far less variability in the change in guessing behavior among subjects for increasingly distant cones as made evident by the more modest estimate $\widehat{\sigma}_2^2 \approx 0.11$. This informs us that the subjects tend to handle increased distances in similar manners in that when the cone distance is lengthened, the subjects tend to increase their guessed distances at similar rates no matter what their initial guess was. The fact that this variance is significantly different from zero simply indicates that the subjects are not identical in their behaviors (those differences can be visually inspected in the figure of the raw response profiles as well).

Lastly, the estimated error variance, $\widehat{\sigma}_\epsilon^2 \approx 3.94$ populates our within-subject covariance matrices. The size of this model-averaged point estimate tells us that our model does not exactly predict the sample data – as our data seem to contain a fair amount of noise – but that the variability in a subject's own responses is likely different from zero, as we might expect. However, this source of variability is certainly not as influential as the variability between subjects.

4 DISCUSSION

4.1 EFFECTS OF MODEL AVERAGING

When we made the decision to use model averaging instead of using the best-by-AICc model, the goal was to improve inferences by accounting for the fact that we had a collection of models deemed important before the analysis of the data began. The effect of our accounting for this additional uncertainty can be seen as a shrinkage effect much like in other shrinkage methods such as LASSO and ridge regression in which unimportant effects are shrunk to zero and allow us to visualize the important covariates' effect on the response, unclouded by effects that are non-zero due to sampling variability and measurement error.

We have seen a portion of said shrinkage effect in the model averaged coefficients presented in Table 3.8 above, but we can investigate the matter further. We have included a collection of plots with one panel per fixed effect in which the vertical direction measures the estimated coefficient for each of the effects. Each of the bubbles represent an estimated coefficient from one of the sixteen *a priori* models and has radius proportional to the model's AICc weight (wider bubbles indicate larger model weights). We have also included 95% unconditional (on model chosen) Wald intervals for each effect. The horizontal bar at the centers of the confidence intervals are the model averaged point estimates for each of the effects.

Figure 4.1 clearly shows the effect of model averaging on the end-result point estimates. Take, for instance, the weight main effect. Weight is included as a non-zero effect in six of the models under consideration and the corresponding coefficient estimates are shown by the non-zero bubbles. We see here that in the models in which weight is included as a non-zero effect, the estimated parameter ranges between about 0.02 and 0.04. Furthermore, for those six models, the model weights are relatively large. For the other nine models, the parameter estimate was estimated at zero and their model weights are small, as depicted by the smaller bubbles overlapping at a vertical position of zero. We then see that the model averaged point estimate really does lie close to the center of

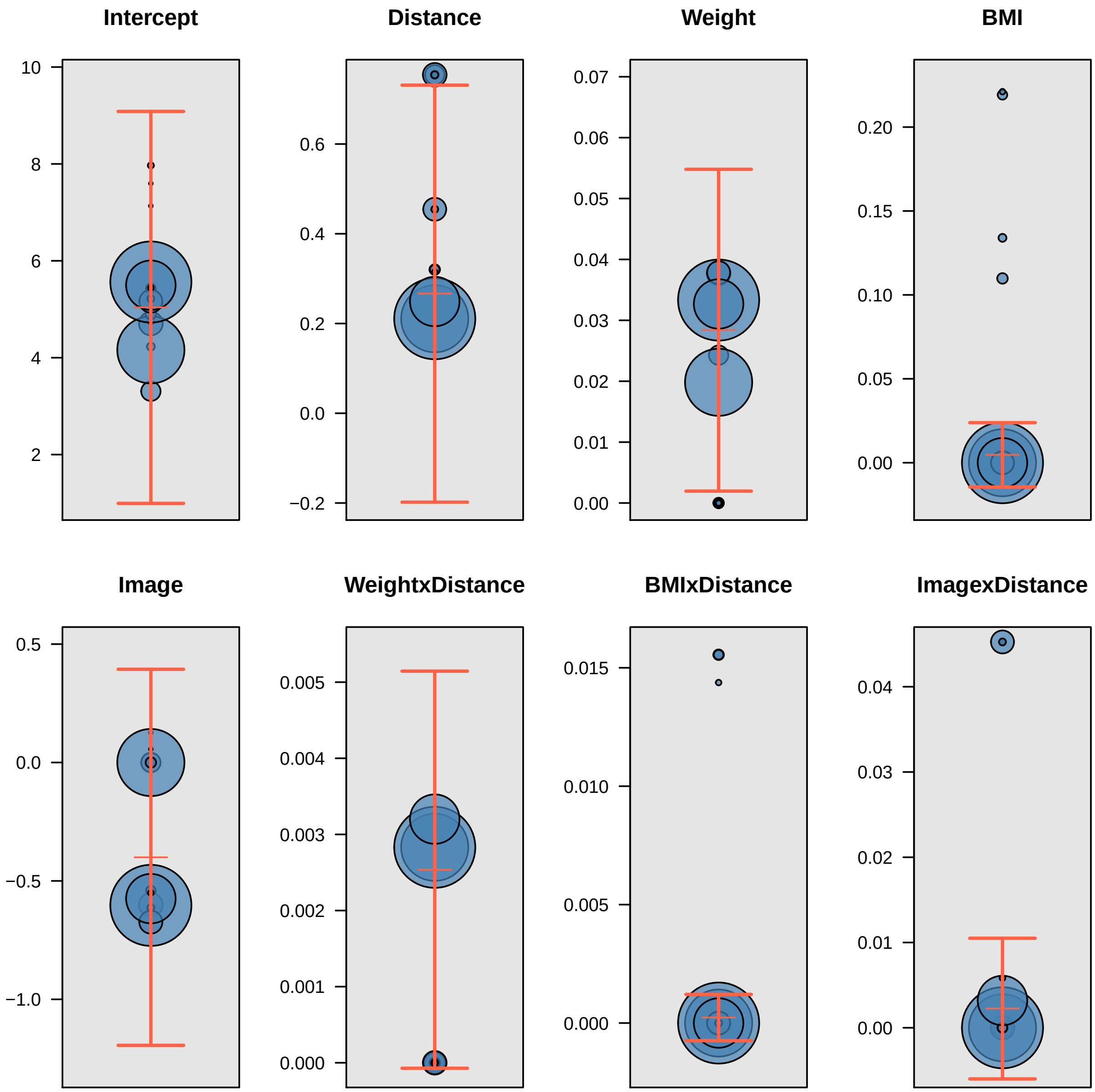


Figure 4.1: Individual fixed effect coefficient estimates from all *a priori* models are illustrated as the vertical position of steel blue bubbles with radius determined by model weight. The light red bars depict 95% confidence intervals unconditional on the model.

the estimates from the top five or so models. There is a perceptible effect of the nine, less AICc important models in that the model averaged point estimate is dragged downwards towards them, if only slightly. The end result is a point estimate that is representative of the estimates from the most important models and the effect is significantly different from zero – notice the confidence interval does not cross zero.

Contrast this with the effect of BMI. Again, BMI was included as a non-zero effect in six of the models and set to zero in the remaining nine. However, the pattern of inclusion for BMI is different from that of weight in that BMI is included in the less important models. In the six models in which BMI is non-zero, the coefficient estimates take on similar roles as weight and are estimated to be between 0.11 and 0.22. However, since those six models are relatively unimportant, the model averaged point estimate is heavily influenced by the more important models in which BMI's effect is set to zero. In the end, BMI's model averaged parameter estimate is very close to zero and is not significantly different from such if we account for model uncertainty.

Next, consider the effect of body image. This effect is taken to be non-zero in ten of the sixteen original models. In nine of those ten models, we estimate an effect of body image to be between about -0.50 and -0.65. If we took any of these models to be the truth and used that model alone in inference, we might estimate that body image is a significant effect. However, the parameter estimates are so widely variable and are included in models with such widely varying importance that its effect is washed out. It's estimated coefficient in the model averaged sense is decreased from the grouping of estimates around -0.60 to about -0.40 and the influence of the variability of estimates due to model averaging causes the effect to be insignificant at the 0.05 level.

Through inspection of the figure, we feel further justified in saying that model averaging is performing just as we expected it to. It allows us to see beyond the 'random noise' introduced by covariate inclusion amongst our models and declare only those effects which are consistently informative to be significant. In our case, those are the main effects for weight and distance and the weight by distance interaction. Evidence from the figure further suggests that there might be a weak body image effect that could be teased out by additional study but we cannot be confident in such a conclusion at this time.

4.2 Relation between Model Averaged and AICc-best Inference

Naturally, since we have taken the route of using model averaging rather than the typical approach of using the best model by AICc to perform inference, we would like to know what we have gained or lost by using the non-standard methods. Theoretically, we would expect that we would be sacrificing precision because we need to account for the additional uncertainty in the model selection process. We would also expect to be able to improve our estimates of effects that are included in the best model and to compute estimates for fixed effects that are not included in the best model.

Below, we have included a plot contrasting the point estimates and 95% confidence intervals for all eight effects using both model averaging and the results from M11, which was our best model overall using AICc. The model averaged effects depicted on the left side of each plot in Figure 4.2 are the same as in Figure 4.1. The estimates from M11 alone are presented on the right of each plot along with corresponding confidence intervals conditional on our choice of M11.

The first item we note is the fact that the model averaged confidence intervals are about the same or wider than those from M11 for all of the effects included in the best model. The reason for the decreased precision is that when we perform model averaging, we are using information from all of the model fits instead of just one. Using the information from the entire collection of models lets us view the effect of each of the covariates on the response from different angles of the parameter space and lets us synthesize all of that information into a single estimate. Since we are combining information about parameters from different models, our confidence intervals are wider than if we had observed the effects from only one point of reference in the parameter space – even though the point of reference for M11 can be considered the best single perspective within our collection of models.

If one likens the technique of model selection to the selection of principal components in a dimension reduction problem, we could make the connection between full model averaging and one's using all of the principal components (and thus retaining the explanatory power of all of the components) and between using the best model alone and using the first principal component only (and thus sacrificing the explanatory power of the rest of the components).

Another important observation is that if we had used M11 only, we would not have been able to say anything about the effects of BMI and the BMI by distance and image by distance interactions

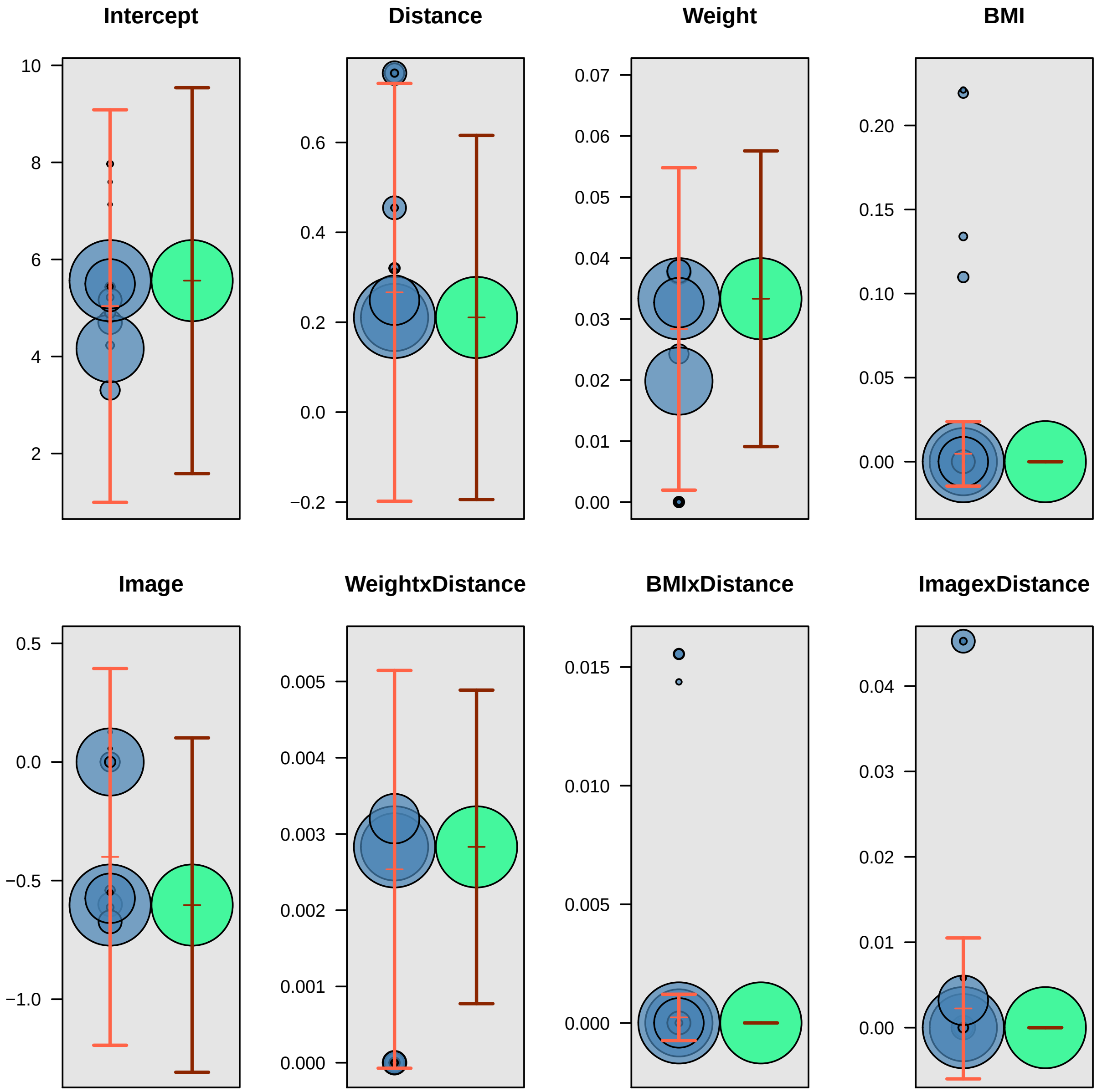


Figure 4.2: Comparison of fixed effects estimates using model averaging (individual model estimates in steel blue and unconditional confidence intervals in light red) on the left and using M11 only (individual model estimates in sea green and conditional confidence intervals in dark red) on the right.

since they were not included in this model. With model averaging, however, we can derive point estimates and make inference for these effects easily. That is, we do not sacrifice such information through the arbitrary choice of any model being best in comparison to the rest.

The size of the effects by model averaging and in M11 are much the same if they are included in M11. We can see this by comparing the midpoints of the confidence intervals to one another.

There are only slight differences between the point estimates with the largest deviations belonging to the main effects for weight and image. The additional information gleaned from observing all of the models allowed us to make the small adjustments to the estimated fixed effects coefficients to better reflect their influence on one's ability to judge distances accurately.

A major source of disagreement between the two approaches is the fact that M11 alone predicts a significant weight by distance interaction effect at the 0.05 significance level (point estimates are 0.0025 from model averaging and 0.0028 from M11). This lends some credence to the effect of differing distance guessing behavior for subjects with different weights and this effect changes among distance guessing occasions. However, once we take into account the model uncertainty, we get a more honest view of this interaction effect and would declare it present but certainly not significant after taking everything into consideration.

Some agreement between can be seen the two approaches in that the effect of body image is non-significant, though there seems to be a negative effect of body image. Another source of agreement lies in those effects not included in M11. Those three effects, despite not being estimable in M11, are estimated as being non-different from zero in the model averaging lens, which gives credence to the choice of M11 being best with respect to including those covariates most important in explaining distance guessing.

4.3 Importance of R^2_{LMM} Values

The overall and partial R^2_{LMM} statistics are less widely used measures of model and covariate worth when compared to information criteria. However, we believe that these statistics allow us to make meaningful conclusions in and of themselves in that they can be used in isolation as non-relative measures of model *and* covariate worth and as relative measures like AICc and others.

From Table 3.4, we can see that the model with distance as the only predictor is best by the R^2_{LMM} measure and M11, the best-by-AICc model, was ranked as sixth among our models. This seems counter-intuitive at first but after some thought about the meaning of R^2_{LMM}, it starts to make some sense after all. What R^2_{LMM} really measures is the ability of the covariate set in each model to explain the multivariate variability of the response. So, the above observation illustrates the fact that distance alone appears to explain about 0.79 of the variability in guessed distances. This is unsurprising as we would ordinarily expect that distance guesses would rely upon the actual distances being judged in some fashion.

If we were to include predictors that explained substantially more variability in the response, then we would expect to see increased R^2_{LMM} values as well. What we see, however, are decreases when we include more of our predictors. The reason for this is that additional predictors' explanatory power does not outweigh their contribution to the estimation of the model covariance components ($\widehat{\sigma}^2_\epsilon$, $\widehat{\sigma}^2_1$ and $\widehat{\sigma}^2_2$) [9]. Take for example the comparison of M2 to M3. When we add the main effect for weight to the model with only the distance fixed effect, we see a decrease in R^2_{LMM} from 0.79 to 0.74 and the estimates for the error variances increase from about 3.95 to about 3.98. The inflation of the error variance estimate seen affects the R^2_{LMM} value and does not exceed the additional explanatory power of the weight variable and so we see a decrease in our goodness of fit statistic.

What we take from this observation is that the distance variable is by far the most important in explaining the response. The important covariates can be ranked: weight, BMI, and image (in that order) and then the two-way interactions with distance, according to R^2_{LMM}.

Our observations are further proliferated by inspection of the partial R^2_{LMM} values. From Table 3.5, we can see that, if we take the average of the partial R^2_{LMM} values across all sixteen models to be meaningful at all, distance is by far the best explainer of distance guesses. Weight and and BMI are the two next most important variables and explain only about 2-4% more of the variability in distance guesses, given the other variables in models in which they appear. We take from this to mean that, of our covariates, weight is the most important but that its effect is small in comparison to distance increases.

4.4 Relation between AICc and R^2_{LMM}

We have used two distinct methods to measure goodness of fit of the models and the variables in which we have been interested. Both the R^2_{LMM} and AICc measures as seen as goodness criteria produced results that are different from one another so much so that we believe that it requires our attention briefly. If we consider model ranking, the order of ranks of models by the two criteria are clearly different. Why might this be so if they are both used to judge goodness of models?

For instance, M11 is first by AICc but sixth by R^2_{LMM} and M2 is first by R^2_{LMM} but seventh by AICc. Also, there does not appear to be any strong patterns to be perceived by looking at the table of ranks alone. Therefore, we appeal to visual clustering of the ranks. Included below is a plot of AICc and R^2_{LMM} values for each of the models in Figure 4.3.

We notice that the relationship between the two measures is not overwhelmingly strong. In fact, the simple Pearson's correlation between the two measures is about -0.17. The direction of the relationship is as expected: as we increase AICc, we tend to see decreases in the R^2_{LMM} values. Said relationship is primarily driven by the three best-by-AICc models, M11, M8, and M15.

The first includes the distance only model, M2, which has a mid-range AICc value and the highest R^2_{LMM}. The second includes the three best-by-AICc models which have comparatively high R^2_{LMM} values. The third grouping includes the rest of the models which have higher AICc values and lower R^2_{LMM} values.

This observation leads us to notice that there are three main groups of models in the plot. In the first group (M2), we see that distance is in a league of its own in terms of explanatory power, but leaves lots of information loss with respect to the rest. In the second group (M8, M11, and M15), when we add the weight and image main effects and their interactions with distance, we decrease information loss (decreasing AICc) but lose a bit of explanatory power due to the low explanatory power embedded in these effects failing to overcome the fluctuations in the variance component estimates. Lastly, the third group (all the rest) consists of all models in which we do not decrease information loss by AICc by adding predictors nor do we explain more of the response in comparison to the model with only distance as a fixed effect.

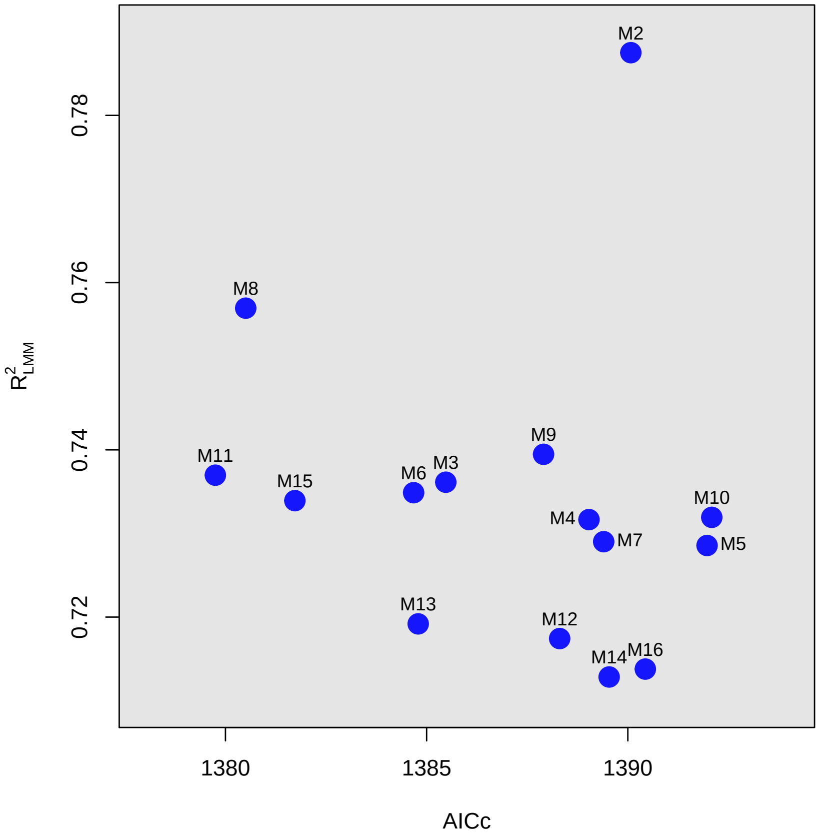


Figure 4.3: R^2_{LMM} versus AICc for each of the sixteen *a priori* models under investigation.

If we consider ranking individual variables rather than models, we can compare the results in Tables 3.5 and 3.8. We see that the $\bar{R}^2_{LMM;parital}$ quantities rank the variables in order of importance: distance then weight, and the rest are competing for explanation of natural variability. The model averaged effects report that the only non-intercept effect that is significantly different from zero is weight. The disagreement here is disconcerting, though it can be defended as in the previous section. We are appreciative for both perspectives because if we had used only the IC approach, we would not have been able to extract the information contained in the fixed effects only and hence observe such in isolation from the information contained in the random portion of the model.

The overall weak correspondence between the two discourages entirely replacing the information-based criterion with the F-test based criterion as a model comparison statistic. However, in conjunction, the two can be used to isolate out the most influential fixed effects and we credit R^2_{LMM} in teasing out the overwhelming effect of distance with respect to the others, whatever we may have thought before analysis.

5 CONCLUSION

We return to the two questions that spurred this investigation.

(I) Does actual body type (as measured by weight or BMI) or perceived body type affect one's ability to accurately judge distances?

(II) Is weight or BMI a better predictor of accurate distance guesses?

Question (I) can be addressed through our results from the model averaged estimates derived from the AICc-based aggregation of the information from the sixteen *a priori* models. Through this lens, we are left with only one of the three main effects of interest that significantly explains guessing behavior: weight. Quantitatively, we expect to see increases in guessed distances of about 2.8 meters – with 95% confidence between 0.2 and 6.0 meters – at any of the distance guessing occasions when weight is increased by 100 pounds, if all other effects are held constant. Also, by Figure 3.8, we visually confirm that the heavier individuals tend to make more accurate guesses of distance compared to their lighter counterparts. This makes sense intuitively as we would expect heavier individuals to be taller and those taller individuals would have greater vantage points from which they may judge distances.

These conclusions also answer question (II). If given a choice between weight and BMI as a predictor for guessing ability, we would urge the reader to choose weight. This comes from our observation of the worth of weight over BMI as an explainer of guessing ability via the R^2_{LMM} statistics and through the AICc model averaged estimates. With the R^2_{LMM} values, weight was declared as the better explainer of guessing behavior given the other variables by explaining about 2% more of the variability in the responses than BMI. Through AICc, we saw that weight and its interaction with distance were the only significant predictors of guessing behavior (other than distance) and that BMI was non-significant due to its exclusion from the most informative of models.

The mean structure (via the fixed effects) are just one aspect of a linear mixed model, the other being the covariance structure. Performing model averaging on the covariance parameter estimates revealed a feature of the data all on its own. Specifically, the sheer magnitude of the variability of subjects' distance guesses for the 10 meter cone masked the effects of some of the other covariates. Practically, this means that since subjects were so different with respect to their distance guessing abilities coming into the study, it became difficult to precisely estimate the influence of weight, BMI, or body image. In addition, the distance guessing behavior across distance guessing occasions do not vary as much as subjects' initial guesses. Indeed, the variability among those behaviors is significantly different from zero and so we conclude that subjects differed significantly in their guessing adjustment behaviors, though this difference is nowhere near the differences in baseline guesses.

From a methodological point of view, the LMM proved to be an effective way to account for the correlation among subjects' own responses since they were measured more than once. The ordinary model placing normal distributions on the random effects and errors seemed to fit well for the most part, though there were some indications that our assumptions of normality and homoskedasticity were not met. We attempted to rectify these two issues but neither of the generalizations that we imposed substantially changed the inferences that we would have made from those models. Since the goal of this application is for the results to be used in a practical, consulting-type setting, we do not believe that the almost indistinguishable model fits substantiated the far more complicated models used to fix those two issues and so we defaulted to the normal-based model for our final inference and conclusions.

Our use of information-based model averaged inference proved to be successful as it allowed us to make decisions regarding the importance of variables that could not populate the same model (weight and BMI) in a comprehensive manner. Instead of basing our conclusions to question (II) on a dichotomous decision between a model containing weight or a model containing BMI, we were able to use a continuous extension of model selection, model averaging, on which we could base our decision. This act greatly expanded the robustness of the decision procedure and allowed us to make conclusions about effects regardless of which models in which they were included.

The coefficient of determination for linear mixed models, R^2_{LMM} and its partial version, proved to be useful in further consolidating the information contained in our sample data. It allowed us to view the information from another perspective and revealed a new feature that might otherwise have been quantitatively (though, perhaps, not intuitively) hidden to us in that changes of distance were the most influential effects on guessing behavior by an order of magnitude over the other effects. It also confirmed our conclusions made using information-based methods and gave us a way to directly judge the ability of predictor sets to explain guessing behavior.

In looking to the future, we would offer some areas of future research in this area to include the practical matters of reducing between-subject variability in guessing behavior at the start of the study through something like a training round or a practice round for the participants to reduce the noise introduced into the data. Also, we might be interested in some other measurements of body image by the participants as other measurements of this characteristic might shed more light on body image's interaction with actual body size. Methodologically, we would recommend expanding the use of the robust linear mixed model to allow for variability that goes beyond normal deviations. Also, we would recommend further investigating the relationship between ML and REML estimation procedures and their interaction among estimation problems and IC-based model ranking.

6 REFERENCES

[1] AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*, 6 (1973), 716–723.

[2] AKAIKE, H. A bayesian extension of the minimum aic procedure of autoregressive model fitting. *Biometrika 66*, 2 (1979), 237–242.

[3] BATES, D. Re: Coefficient of determination (r2) when using lme(). http://thread.gmane.org/gmane.comp.lang.r.lme4.devel/684. Accessed: 1/14/2015.

[4] BOX, G., AND DRAPER, N. *Empirical model-building and response surfaces*. Wiley, 1987.

[5] BURNHAM, K., AND ANDERSON, D. Model selection and multimodel inference. *Sociological Methods and Research 33*, 2 (2004), 261–304.

[6] BURNHAM, K., ANDERSON, D., AND HUYVAERT, K. Aic model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behavioral Ecology and Sociobiology 65* (2011), 23–35.

[7] CASELLA, G., AND BERGER, R. *Statistical Inference*, 2 ed. Cengage Learning, 2002.

[8] CHEN, X., ZOU, G., AND ZHANG, X. Frequentist model averaging for linear mixed effects models. *Frontiers of Mathematics in China 8*, 3 (2013), 497–515.

[9] EDWARDS, L., MULLER, K., WOLFINGER, R., QAQISH, B., AND SCHABENBERGER, O. An r2 statistic for fixed effects in the linear mixed model. *Statistics in Medicine 27* (2008), 6137–6157.

[10] FANG, Y. Asymptotic equivalence between cross-validations and akaike information criteria in mixed-effects models. *Journal of Data Science 9* (2011), 15–21.

[11] FITZMAURICE, G., LAIRD, N., AND WARE, J. *Applied longitudinal analysis*, 2nd ed. John Wiley & Sons, 2011.

[12] GIVENS, G., AND HOETING, J. *Computational Statistics*, 2nd ed. Wiley, 2012.

[13] GURKA, M. Selecting the best linear mixed model under reml. *The American Statistician 60*, 1 (2006), 19–26.

[14] HARVILLE, D. Bayesian inference for variance components using only error contrasts. *Biometrika 61*, 2 (1974), 383–385.

[15] HARVILLE, D. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association 72*, 358 (1977), 320–338.

[16] HOETING, J., MADIGAN, D., RAFTERY, A., AND VOLINSKY, C. Bayesian model averaging: A tutorial (with discussion). *Statistical Science 14* (1999), 382–417.

[17] JOHNSON, R., AND WICHERN, D. *Applied Multivariate Statistical Analysis*, 6th ed. Pearson, 2007.

[18] KUTNER, M., NACHTSHEIM, C., AND NETER, J. *Applied Linear Regression Models*. McGraw-Hill/Irwin, 2004.

[19] LAIRD, N., AND WARE, J. Random-effects models for longitudinal data. *Biometrics 38*, 4 (1982), 963–974.

[20] LITTELL, R., MILLIKEN, G., STROUP, W., WOLFINGER, R., AND SCHABENBERGER, O. *SAS for mixed models*, 2 ed. SAS Institute, Inc., 2006.

[21] MCCULLOCH, C., AND SEARLE, S. *Generalized, linear, and mixed models*. John Wiley & Sons, 2001.

[22] MCCULLOCH, J. On heteros*edasticity. *Econometrica 53*, 2 (1983), 483.

[23] NAKAGAWA, S., AND SHIELZETH, H. A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution 4* (2013), 133–142.

[24] PINHEIRO, J., LIU, C., AND WU, Y. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics 10*, 2 (2001), 249–276.

[25] RAO, C. A note on a generalized inverse of a matrix with applications to problems in mathematical statistics. *Journal of the Royal Statistical Society Series B 24*, 1 (1962), 152–158.

[26] SAS INSTITUTE, INC. *SAS/STAT(R) 9.2 User's Guide, Second Edition*, 2009. http://support.sas.com/documentation/cdl/en/statug/63033/PDF/default/statug.pdf.

[27] SINGER, J., NOBRE, J., AND ROCHA, F. Diagnostic and treatment for linear mixed models. *Proceedings 59th ISI World Statistics Congress Session CPS203* (August 2013), 5486–5491.

[28] STONE, M. An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society Series B (Methodological) 39*, 1 (1977), 44–47.

[29] SYMONDS, M., AND MOUSSALLI, A. A brief guide to model selection, multimodel inference, and model averaging in behavioral ecology using akaike's information criterion. *Behavioral Ecology and Sociobiology 65* (2011), 13–21.

[30] TAYLOR, J., WITT, J., AND SUGOVIC, M. When walls are no longer barriers: Perception of wall height in parkour. *Perception*, 40 (2011), 757–760.

[31] VAN DYK, D. Fitting mixed-effects models using efficient em-type algorithms. *Journal of Computational and Graphical Statistics 9*, 1 (2000), 78–98.

[32] VERBEKE, G., AND MOLENBERGHS, G. *Linear mixed models for longitudinal data*, 1st ed. Springer, 1997.

[33] WITT, J., BAKDASH, J., AUGUSTYN, J., COOK, A., AND PROFFITT, D. The long road of pain: Chronic pain increases perceived distance. *Experimental Brain Research 192* (2009), 145–148.

[34] WITT, J., AND PROFFITT, D. See the ball, hit the ball: Apparent ball size is correlated with batting average. *Psychological Science*, 16 (2005), 937–938.

[35] WOLFINGER, R. Covariance structure selection in general mixed models. *Communications in Statistics – Simulation and Computation 22* (1993), 1079–1106.

[36] ZHU, M., AND CHIPMAN, H. Darwinian evolution in parallel universes: A parallel genetic algorithm for variable selection. *Technometrics 48*, 4 (2006), 491–502.