Prediction in Multilevel Models[1]

David Afshartous
School of Business Administration, University of Miami,
Coral Gables, FL 33124-8237
Jan de Leeuw
Department of Statistics, University of California,
Los Angeles, CA 90095-1554

ABSTRACT: Multilevel modeling is an increasingly popular technique for analyzing hierarchical data. We consider the problem of predicting a future observable y_{*j} in the jth group of a hierarchical dataset. Three prediction rules are presented and assessed via a Monte Carlo study that extensively covers both the sample size and parameter space. Specifically, the sample size space concerns the various combinations of level level-1 and level-2 sample sizes, while the parameter space concerns different intraclass correlation values. The three prediction rules employ OLS, Prior, and Multilevel estimators for the level-1 coefficients β_j. The multilevel prediction rule performs the best across all design conditions, and the prior prediction rule degrades as the number of groups J increases.

KEY WORDS: prediction, Monte Carlo, multilevel model

1 Introduction

Prediction in multilevel models is considered in terms of forecasting unobserved (yet observable) units at the individual level. Consider the school example. After carrying out a multilevel model analysis on some data, suppose we want to know the outcome (y) for a student not in the data set. Formally, let y_{*j} be the unknown outcome measure, say mathematics score, for an unsampled student in the jth school, where school j is not necessarily in our sample or even known. The basic problem is to predict y_{*j}. We present three main approaches to the prediction of y_{*j} and examine their performance through a simulation study that extensively covers both the sample size and parameter space. In addition, we compare these results with the corresponding results for estimation.

Although there exists an extensive literature on estimation issues in multilevel models, the same cannot be said with respect to prediction. Exceptions include Rubin's seminal Law School Validity Studies paper where a multilevel model without group level covariates is used to predict first year GPA based on LSAT score; he found that predictions were improved via what he termed Empirical Bayes predictors. Gray et al.(2001) consider the problem of predicting future 'value-added' performance across groups from past trends. The main result is that such prediction is unreliable.[2] However, there does not exist a full treatment of the multilevel prediction problem. The multilevel prediction is an important

[1]This research was supported by a grant from the National Institute for Statistical Sciences

[2]They examined A/AS level results obtained by English institutions from year to year. Their approach is different from our approach as they are considering cohort periods and are not predicting y_{*j}.

problem given the popularity of multilevel models in a variety of fields and the usefulness of being able to forecast future observations.

In section 1.1 we review the multilevel model and in section 1.2 we discuss estimation in multilevel models. In section 2 we present three approaches to prediction in multilevel models, and in section 3 we describe the simulation study design with which we assess these three methods. Results and discussion are in section 4 and a brief summary is in section 5.

1.1 The Multilevel Model

Multilevel modeling is a statistical technique designed to facilitate inferences from hierarchical data. Other names such as hierarchical linear modeling, random coefficient modeling, or Empirical Bayes estimation, are often employed, usually as a function of one's research discipline. Nevertheless, the basic framework is the same in each case: a given data point y_{ij} represents the ith observation in the jth group, e.g., the ith student in the jth school for educational data; we may have J groups, where the jth group contains n_j observations. Although several levels of data may be considered, this discussion is restricted to the simple case of primary units grouped within secondary units and we will periodically refer to the applied example of students (level-1) grouped within schools (level-2). Within each group, we have the following level-1 model equation:

$$Y_j = X_j\beta_j + r_j \tag{1}$$

Each X_j has dimensions $n_j \times p$, and $r_j \sim N(0, \sigma^2\Psi_j)$, with Ψ_j usually taken as I_{n_j}. To be sure, these J regression equations may be estimated separately, thereby ignoring the structure in the data. A common problem with this approach, however, is that some of the groups do not contain sufficient data to produce stable estimates. In multilevel modeling, this problem is remedied by modeling some or all of the level-1 coefficients, β_j, as random variables.[3] They may also be functions of level-2 (school) variables:

$$\beta_j = W_j\gamma + u_j \tag{2}$$

Each W_j has dimension $p \times q$ and is a matrix of background variables on the jth group and $u_j \sim N(0, \tau)$. Clearly, since τ is not necessarily diagonal, the elements of the random vector β_j are not independent. For instance, there might exist a covariance between the slope and intercept for each regression equation. Equation 2 may be viewed as a prior for the distribution of the level-1 β_j, modeled as varying around a conditional grand mean $W_j\gamma$ with a common variance τ, thereby expressing a judgment of similarity with respect to the groups.[4] For instance, in the school example, this expresses the reasonable judgment that schools, although unique in many ways, have certain common characteristics that may be accounted for in the modeling process. Furthermore, the separate equations for level-1 and level-2 data readily models/displays the relationship between variables from different levels of the data, where the magnitude of the elements of γ measure the strength of these cross-level interactions. Specifically, the group level-2 variables may either increase or decrease the

[3] Viewing equation 1 as a model which describes a hypothetical sequence of replications which generated the data, the introduction of random coefficients expresses the idea that the intercepts and slopes are no longer fixed numbers—which are constant within schools and possibly between schools—and that they may vary over replications (de Leeuw & Kreft, 1995).

[4] Thus, given an estimate $\hat{\gamma}$ the prior estimate for β_j would be $W_j\hat{\gamma}$.

individual level-1 coefficients. For the school example these phenomena would be classified as "school effects."

Combining equations yields the single equation model:

$$Y_j = X_j W_j \gamma + X_j u_j + r_j, \tag{3}$$

which may be viewed as a special case of the mixed linear model, with fixed effects γ and random effects u_j.[5] Researchers more interested in the fixed effects γ rather than the level-1 coefficients β_j often prefer this formulation. Marginally, y_j has expected value $X_j W_j \gamma$ and dispersion $V_j = X_j \tau {X_j}' + \sigma^2 I$. Observations in the same group have correlated disturbances, and this correlation will be larger if their predictor profiles are more alike in the metric τ (De Leeuw & Kreft 1995). Thus, the full log-likelihood for the jth unit is

$$L_j(\sigma^2, \tau, \gamma) = -\frac{n_j}{2}\log(2\pi) - \frac{1}{2}\log|V_j| - \frac{1}{2} d_j' V_j^{-1} d_j, \tag{4}$$

where $d_j = Y_j - X_j W_j \gamma$. Since the J units are independent, we write the log-likelihood for the entire model as a sum of unit log-likelihoods, *i.e.*,

$$L(\sigma^2, \tau, \gamma) = \sum_{j=1}^{J} L_j(\sigma^2, \tau, \gamma). \tag{5}$$

By appropriately stacking the data for each of the J level-2 units, we may write the model for the entire data without subscripts. Thus, we have:

$$Y = X\beta + r \tag{6}$$

with r normally distributed with mean 0 and dispersion Ψ where

$$\begin{aligned} Y &= (Y_1', Y_2', \ldots, Y_J')', \\ \beta &= (\beta_1', \beta_2', \ldots, \beta_J')', \\ r &= (r_1', r_1', \ldots, r_J')' \end{aligned}$$

and

$$X = \begin{pmatrix} X_1 & 0 & \ldots & 0 \\ 0 & X_2 & \ldots & 0 \\ . & . & \ldots & . \\ . & . & \ldots & . \\ . & . & \ldots & . \\ 0 & 0 & \ldots & X_J \end{pmatrix} \quad \Psi = \begin{pmatrix} \Psi_1 & 0 & \ldots & 0 \\ 0 & \Psi_2 & \ldots & 0 \\ . & . & \ldots & . \\ . & . & \ldots & . \\ . & . & \ldots & . \\ 0 & 0 & \ldots & \Psi_J \end{pmatrix}$$

We may also write the level-2 equation in no-subscript form through similar stacking manipulations:

$$\beta = W\gamma + u \tag{7}$$

[5]For an excellent review of estimation of fixed and random effects in the general mixed model see Robinson, 1991.

where u is normally distributed with mean 0 and covariance matrix T where

$$\begin{aligned} W &= (W_1', W_2', \ldots, W_J')', \\ u &= (u_1', u_2', \ldots, u_J')', \\ T &= \begin{pmatrix} \tau & 0 & \ldots & 0 \\ 0 & \tau & \ldots & 0 \\ . & . & \ldots & . \\ . & . & \ldots & . \\ . & . & \ldots & . \\ 0 & 0 & \ldots & \tau \end{pmatrix} \end{aligned}$$

Combining equations, the entire model may be written as:

$$Y = XW\gamma + Xu + r \tag{8}$$

where we note that $E(y) = XW\gamma$ and $\text{Var}(y) = XTX' + \Psi$.

1.2 Estimation

Given that the multilevel model may be viewed from a variety of perspectives (e.g., separate equation model versus combined equation model), so can the approaches to estimation. Raudenbush & Bryk (2002) discuss estimation in multilevel models by casting the multilevel model as a particular case of the general Bayes linear model and hence present estimates of β_j as posterior means of their corresponding posterior distribution. Other approaches focus on the James-Stein "borrowing-of-strength" aspect of multilevel modeling when presenting estimates of the level-1 coefficients.[6] Another alternative is to focus on the likelihood established by equation 5, where maximum likelihood estimates for the three parameters σ^2, τ, and γ are obtained. Regardless, the main result is that the estimates of β_j may be expressed as a linear combinations of the OLS estimate $\hat{\beta}_j = ({X'}_j X_j)^{-1} X_j y_j$ and—given an estimate of γ—the prior estimate $W_j\hat{\gamma}$ of β_j, the weights being proportional to the estimation variance in the OLS estimate and the prior variance of the distribution of β_j. Thus, this may be viewed as a compromise between the within-group estimator which ignores the data structure and the between-group estimator which models he within-group coefficients as varying around a conditional grand mean. More formally, assuming for now that the variance components and γ are known, the multilevel model estimate of β_j may be expressed as follows:

$$\hat{\beta}_j^{\,*} = \Theta_j \hat{\beta}_j + (I - \Theta_j) W_j \gamma \tag{9}$$

where

$$\Theta_j = \tau(\tau + \sigma^2 ({X_j}' X_j)^{-1})^{-1} \tag{10}$$

is the ratio of the parameter variance for β_j (τ) relative to the variance $\sigma^2({X_j}' X_j)^{-1}$ for the OLS estimator for β_j plus this parameter variance matrix. Thus, if the OLS estimate is

[6] Recall that since the level-1 coefficient β_j is a random variable, the term "estimation" is being employed somewhat pejoratively here.

unreliable, $\hat{\beta_j}^*$ will pull $\hat{\beta}_j$ towards $W_j\hat{\gamma}$, the prior estimate.[7] Indeed, a little bit of algebra demonstrates that the shrinkage estimator in equation 9 is the expected value of β_j given y_j:[8]

$$\begin{aligned} E(\beta_j|y_j) &= E(\beta_j) + \mathrm{Cov}(\beta_j, y_j)(\mathrm{Var}(y_j))^{-1}(y_j - E(y_j)) \\ &= W_j\gamma + \tau {X_j}' {V_j}^{-1}(y_j - X_j W_j \gamma) \\ &= W_j\gamma + \tau {X_j}' {V_j}^{-1} y_j - \tau {X_j}' {V_j}^{-1} X_j W_j \gamma \end{aligned} \tag{11}$$

Swamy (1971, p.101) presents the following formula for the inverse of V_j,

$$ {V_j}^{-1} = \sigma^2[X_j({X_j}'X_j)^{-1}{X_j}'] + X_j({X_j}'X_j)^{-1}{A_j}^{-1}({X_j}'X_j)^{-1}{X_j}' \tag{12}$$

where $A_j = \tau + \sigma^2({X_j}'X_j)^{-1}$. This implies that ${X_j}'{V_j}^{-1}X_j = {A_j}^{-1}$ and that ${X_j}'{V_j}^{-1}y_j = {A_j}^{-1}\hat{\beta}_j$ (de Leeuw & Kreft 1986). Substituting these two results into the previous equation quickly leads to the desired result:

$$\begin{aligned} E(\beta_j|y_j) &= W_j\gamma + \tau {A_j}^{-1}\hat{\beta}_j - \tau {A_j}^{-1} W_j \gamma \\ &= \tau {A_j}^{-1}\hat{\beta}_j + (I - \tau {A_j}^{-1}) W_j \gamma \\ &= \Theta_j \hat{\beta}_j + (I - \Theta_j) W_j \gamma \end{aligned}$$

The conditional expectation representation of the shrinkage estimator is well known as the minimum mean square linear estimator (MMSLE) of β_j (Chipman 1964, Rao 1965b).[9]

One may also write the multilevel estimate as $\hat{\beta_j}^* = W_j\gamma + \hat{u_j}$, where we recall that u_j may be interpreted in the mixed model sense as the random effect of the jth group. From the literature on the estimation of random effects in mixed linear models, we have the commonly employed estimator of random effects:

$$\hat{u}_j = {C_j}^{-1}{X_j}'(y_j - X_j W_j \gamma) \tag{13}$$

where

$$C_j = {X_j}'X_j + \sigma^2\tau \tag{14}$$

To be sure, the fixed effects γ are usually unknown and must be estimated. The estimation of the fixed effects is most easily discussed by ignoring the level-1 β_j's altogether. In doing so, one focuses instead on the combined equation 3, where the problem then becomes one of

[7] The shrinkage estimator in equation 9 is often referred to as a Bayes or posterior estimator.

[8] Recall that we have y_j and β_j distributed multivariate normal with $E(y_j) = X_j W_j \gamma$, $E(\beta_j) = W_j\gamma$ and $\mathrm{Cov}(\beta_j, y_j) = \mathrm{Cov}(\beta_j, X_j\beta_j + r_j) = \mathrm{Cov}(\beta_j, X_j\beta_j) = \tau {X_j}'$. And, employing the well known result that the conditional expectation in the normal case is equivalent to the linear regression of β_j on y_j leads to the result in equation 11.

[9] Note: since we are "estimating" a random variable, care must be taken with respect to notation. Given an observed random variable y and an unobservable random variable w, let $t(y)$ be an estimator of the realized value of the random variable w. The MSE of $t(y)$ is defined as $E(t(y) - w)^2$, where all expectations are taken with respect to the joint distribution of y and w. We say that $t(y)$ is unbiased if $E(t(y)) = E(w)$. Given that the prediction error of $t(y)$ equals $t(y) - w$, we have that $t(y)$ unbiased implies that the MSE of $t(y)$ equals the variance of its prediction error.

estimating the fixed effects γ in a mixed linear model, the result of which is the well-known formula:

$$\hat{\gamma} = (\sum_{j=1}^{J} W_j' X_j' V_j^{-1} X_j W_j)^{-1} \sum_{j=1}^{J} W_j' X_j' V_j^{-1} y_j \tag{15}$$

where

$$V_j = \mathrm{Var}(y_j) = X_j \tau X_j' + \sigma^2 I$$

One may interpret the above estimator of γ as a generalized linear model (GLM) estimator. In the case of unknown γ, the shrinkage estimator of equation 9 employing this estimator of γ yields the minimum mean square linear unbiased estimator (MMSLUE) of β_j (Harville 1976).[10] de Leeuw & Kreft (1995) discuss alternative estimates of the fixed effects via a two-step procedure, where one first obtains the OLS estimates of the β_j and then regresses these values on the W_j values. Regardless, this approach of focusing on the estimation of γ instead of β_j is preferred by some since we are actually estimating a parameter and do not may blur the distinction that β_j is a random variable. Furthermore, casting the multilevel model in the mixed model framework links multilevel model prediction to the more natural prediction problems that occur in such areas as repeated measures studies (See Rao 1987).

The prior discussion assumes that the variance components are known. Although there is considerable agreement with respect to the estimation of fixed effects, there is significantly less agreement with respect to the variance components. The maximum likelihood estimates of the variance components must be computed iteratively, via procedures such as Fisher Scoring (Longford, 1987), iteratively reweighted generalized least squares (Goldstein, 1986), or the EM algorithm (Dempster, Laird, & Rubin, 1977). These and other procedures manifest themselves in several software packages: HLM (Raudenbush et al., 2000), MIXOR (Hedeker & Gibbons, 1996), MLWIN (Rabash et al., 2000), SAS Proc Mixed (Littell et al., 1996), and VARCL (Longford, 1988). In addition, the software package BUGS (Spriegelhalter et al., 1994) incorporates fully Bayesian methods that have been introduced (Gelfant et al., 1990; Seltzer, 1993). Note, although Lindley & Smith (1972) provided a general framework for hierarchical data with complex error structures, the inability to estimate the covariance components for unbalanced data precluded using such models in practice. The introduction of the EM algorithm provided a numeric solution to this problem and paved the way to various other approaches mentioned above.

Although estimation in multilevel models is an important topic, it is not the focus of this paper. The focus here lies in the prediction of a future observable y_{*j} and will be elaborated in the next section.

2 Prediction in Multilevel Models

Prediction in multilevel models is considered in terms of forecasting unobserved (yet observable) units, either at level-1 or level-2. A concise definition is important, for the potential for confusion arises from the close link of the multilevel model with the mixed linear model

[10] One must restrict oneself to the class of unbiased estimators since a MMSLE does not exist for the unknown γ case (Pfefferman 1984).

where one finds the term "prediction" reserved for estimating/predicting random effects.[11] Consider the school example. After carrying out a multilevel model analysis on some data, suppose we want to know the outcome (y) for a student not in the data set. Formally, let y_{*j} be the unknown outcome measure, say mathematics score, for an unsampled student in the jth school, where school j is not necessarily in our sample or even known. Furthermore, let us assume that the multilevel model structure given above is true, although we know that the model is never true. The basic problem is to predict y_{*j}. We present three main approaches to the prediction of y_{*j} and examine their performance through a simulation study that extensively covers both the sample size and parameter space.

The three methods that will be examined are multilevel prediction, prior prediction, and OLS prediction. These three predictive methods correspond to the three possible ways of estimating β_j for multilevel data discussed previously. The relative properties of these estimators is not of central interest, for the focus is on the prediction of a future observable—estimation is a means to an end.[12] However, whether or not the results herein agree with multilevel studies on estimation is of interest. Guidelines exist for appropriately choosing the level-1 and level-2 sample sizes exist with respect to the estimation of fixed effects and variance components. (Busing 1993, Bassiri 1988, Kim 1990, Mok 1995)

2.1 OLS Prediction Method

In this approach we emphasize that there is no level-2 model, i.e., the level-1 β_j coefficients are not modeled as random variables regressed on level-2 variables. Instead, there are simply J separate regression equations:

$$Y_j = X_j\beta_j + r_j, \tag{16}$$

and, as before, the goal is to predict a future observation in the jth group, y_{*j}:

$$y_{*j} = X_{*j}\beta_j + r_{*j} \tag{17}$$

If y_{*j} were observed X_{*j} would merely represent a row of the X_j design matrix and that $r_{*j} \sim N(0, \sigma^2)$. For the prediction of y_{*j} one simply takes the OLS estimate estimate $\hat{\beta}_j$ obtained solely from the jth group and employs the following prediction rule:

$$\hat{y}_{*j} = X_{*j}\hat{\beta}_j \tag{18}$$

where

$$\hat{\beta}_j = (X_j{}'X_j)^{-1}X_j y_j$$

Thus, in spite of the nested nature of the data and the fact that the assumption of a diagonal dispersion matrix is violated (recall that $V_j = X_j\tau X_j + \sigma^2 I$), the conventional OLS procedure is utilized. There exists the risk of unstable prediction in the cases where the number of units within the groups is small and over-fitting is a common problem for OLS. Nevertheless, there is the positive benefit of using a well-known and more easily communicable statistical procedure.

[11] Some authors rebel strong against the term prediction since the random effects under investigation may have occurred thousands of years ago.

[12] To be sure, in the school example neither the student nor the school official is concerned about coefficients estimates; rather, the focus is on the outcome and the more accurate we can predict the outcome the better.

2.2 Prior Prediction Method

In this case, the structure of the data is not ignored; instead, the setup of the multilevel model is adopted. However, we stop short of an actual multilevel analysis, treating the level-2 model equation as a prior for β_j and employing the estimate of that prior as our estimate for β_j. The technique for estimating γ shall be that which was presented in equation 15. Recalling that the multilevel estimate can be viewed as a weighted combination of the OLS estimate and the prior estimate, this approach corresponds to putting all of the weight on the prior. Hence, the prediction rule now becomes:

$$\hat{y}_{*j} = X_{*j} W_j \hat{\gamma} \tag{19}$$

where

$$\hat{\gamma} = (\sum_{j=1}^{J} W_j' X_j' \hat{V}_j^{-1} X_j W_j)^{-1} \sum_{j=1}^{J} W_j' X_j' \hat{V}_j^{-1} y_j \tag{20}$$

and

$$\hat{V}_j \;\; = \;\; \hat{\mathrm{Var}}(y_j) = X_j \hat{\tau} X_j' + \hat{\sigma}^2 I$$

where $\hat{\tau}$ and $\hat{\sigma}^2$ must be estimated iteratively via maximum likelihood. Consider the case when we do not have any level-2 W_j information. In such a case one may view all of the β_j as randomly varying around some mean level γ. Then in the prediction above the estimate of this mean level would be substituted for the estimate of each β_j. By introducing the level-2 W_j information the concept of conditional exchangeability is being modeled, i.e., given two schools with the same W_j information one expects their β_j to vary around the same mean level. To be sure, basing predictions on a conditional grand mean will produce a much different prediction than that produced from the OLS method. However, it does utilize the entire data and will thus not be vulnerable to small sample instability problems.

The prior prediction method may be viewed as a diagnostic check of the multilevel under consideration. Recall, the multilevel model is often used in an attempt to "borrow strength" in the James-Stein sense. Groups are modeled as conditionally exchangeable and estimates are formed as weighted combinations of an ensemble estimate and a solo estimate, the ensemble being the prior and the solo being the OLS. If the multilevel model under consideration is poor or incorrect, the "borrowing" of strength will not be a good idea, i.e., the estimate should not be pulled toward the ensemble estimate, and neither should any prediction. Hilden-Minton (1995) discusses this with respect to diagnostics and further developed Geisser's (1979) model criticism for the multilevel model.

2.3 Multilevel Prediction Method

In this case the prediction rule is formed using the multilevel model estimate of β_j. Recall that this estimate may be written as follows:

$$\hat{\beta_j}^{*} = \Theta_j \hat{\beta}_j + (I - \Theta_j) W_j \hat{\gamma} \tag{21}$$

where

$$\Theta_j = \tau(\tau + \sigma^2 (X_j' X_j)^{-1})^{-1} \tag{22}$$

is the ratio of the parameter variance for β_j (τ) relative to the variance for the OLS estimator for β_j plus this parameter variance matrix. Thus, if the OLS estimate is unreliable, $\hat{\beta_j}^*$ will pull $\hat{\beta_j}$ towards $W_j\hat{\gamma}$, the prior estimate. With regard to the prediction rule, the multilevel estimate β_j is used to form the multilevel predictor:

$$\hat{y}_{*j} = X_{*j}\hat{\beta_j}^* \tag{23}$$

Given that the multilevel model estimate of the level-1 coefficient is a shrinkage estimator, much of the multilevel literature revolves around the advantage of shrinkage estimators, how they borrow strength and solve the instability of estimation problem along along with many other issues encountered when dealing with nested data. With respect to prediction, Goldstein et al.(2001) consider the problem of predicting future 'value-added' performance across groups from past trends. The main result is that such prediction is unreliable.[13] Rubin (1980) examined the performance of what he termed empirical Bayes predictors in his Law School Validity Studies paper. His approach is similar to our approach of focusing on the predicting the future observable y_{*j}. However, his empirical Bayes predictor can be viewed as the basic multilevel model without any level-2 variables. For his particular data set he showed small gains using empirical Bayes predictors. However, he did not employ any level-2 variables to extend his model to a full multilevel model. His searches for useful level-2 variables to improve prediction failed to produce any viable candidates.[14] In addition to the advantage of shrinkage estimators, there is also a small literature on the dangers of shrinkage estimators, giving rise to limited translation rules. (Efron & Morris 1971, 1972) These represent safeguards for shrinking the estimators too far toward the ensemble estimate. The same concern exists for prediction, where predictions may be translated or shrunk too far, resulting in various practical worries.[15]

The relative performance of the three prediction rules is assessed via an extensive simulation study which is described in the next section.

3 Simulation Study Design

Multilevel data is simulated under a variety of design conditions, closely following the simulation study of Busing (1993) where the distribution of level-2 variance component estimates was examined. As in Busing (1993), a simple 2-level multilevel model with one explanatory variable at each level and equal numbers of units per group is considered. A two-stage simulation scheme is employed. At the first stage the level-1 random coefficients are generated according to the following equations:

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01}W_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}W_j + u_{1j} \end{aligned}$$

[13] They examined A/AS level results obtained by English institutions from year to year. Their approach is different from our approach as they are considering cohort periods and are not predicting y_{*j}.

[14] Personal communication with D. Rubin, 7/97.

[15] In Rubin's Law School research the law school officials would be concerned about predictions that are translated too far in the positive direction, while the applicants would be worried about their individual predictions being translated too far in the negative direction.

The γ's are the fixed effects and are set to a predetermined value; they are set all equal to one as in Busing (1993). W_j is a standard normal random variable, while the error components, u_{0j} and u_{1j}, have a bivariate normal distribution with mean $(0, 0)$ and a 2×2 covariance matrix τ. The two diagonal elements of τ, τ_{00} and τ_{11}, are equal in each design condition. The off-diagonal covariance term τ_{01} will then determine the correlation between the intercept and slope:

$$r_{u_{0j},u_{1j}} = \frac{\tau_{01}}{(\tau_{00}\tau_{11})^{1/2}} \tag{24}$$

Another parameter of interest in the simulation design is the intraclass correlation ρ. The intraclass correlation is defined as follows:

$$\rho = \frac{\tau_{00}}{\tau_{00} + \sigma^2} \tag{25}$$

and thus measures the degree to which units within the same unit are related. Intraclass correlations of 0.2 and above are common in educational research; a range of intraclass values of 0.2, 0.4, 0.6, and 0.8 is examined in order to provide information for both high and low intraclass correlation conditions.

The second stage of the simulation concerns the first level of the multilevel model, where observations are generated according to the following equation:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \epsilon_{ij} \tag{26}$$

The level-2 outcome variables, the β's, were determined at the first stage of the simulation. The level-1 explanatory variable, X_{ij}, is simulated as a standard normal random variable, while the level-1 error ϵ_{ij} is a normal random variable with mean 0 and variance σ^2 specified as .5. Since only the balanced data case is considered, where there are n units grouped within J groups, a total of Jn outcomes are simulated. In order to study prediction, an extra $(n + 1)$st observation is simulated for each of the J groups; this observation is set aside and is not used for estimative purposes; this is the future observable y_{*j} for which the prediction rules are applied. Table 1 and Table 2 summarize the various parameter specifications in the simulation design.

Intra-class correlation ρ	0.200	0.400	0.600	0.800
Variance τ_{00}, τ_{11}	0.125	0.333	0.750	2.00

Table 1: ρ, τ_{00}, τ_{11}

	Correlation intercepts-slopes		
Variance	0.25000	0.5000	0.75000
0.125	0.03125	0.0625	0.09375
0.333	0.08330	0.1667	0.25000
0.75	0.18750	0.3750	0.56250
2.0	0.50000	1.0000	1.50000

Table 2: τ_{01}

Simulations are conducted under various sample size combinations for the number of groups (J) and the number of observations per group (n). Information concerning the effects of J and n with respect to the performance of prediction rules is of practical interest at the design or data gathering phase. To be sure, given one's research interests, one would want to know the appropriate values for the number of groups and number of elements per group to sample, especially given the increased cost of including an additional group in one's study. Thus, an extensive sample size space is explored in this simulation study. The layout of the design is given in Table 3:

	n_j				
J	5	10	25	50	100
10	50	100	250	500	1000
25	125	250	625	1250	2500
50	250	500	1250	2500	5000
100	500	1000	2500	5000	10000
300	1000	3000	7500	15000	30000

Table 3: Sample sizes

Each design specification depends on the level of the parameters and the $J \times n$ sample sizes. There are twenty-five possible $J \times n$ combinations and twelve possible parameter specifications, yielding a total of 300 design conditions. As mentioned above, one additional observation per group is simulated which is used to assess the prediction rules. Thus, when $J = 10$ there will be 10 predictions for a given dataset. In addition, for each design condition 100 replications are performed, i.e., 100 multilevel data sets are simulated for each design condition and prediction is assessed within each of these replications. Thus, since there are 300 design conditions, a total of $30,000$ multilevel data sets will be generated in this initial part of the study.

This next phase of this simulation study represents a comparison of the three predictors presented earlier: multilevel, prior, and OLS. Recall that the goal is to predict a future observable y_{*j} in each of our J groups and replicate this process 100 times to account for variability. The adequacy of prediction is measured via predictive mean square error (PMSE), where the popular technique of taking the average of the sum the squared errors (SSE) of the observed and predicted values is employed.[16] Thus, for each of the 300 design conditions there are 100 replications of the predictive mean square error for each prediction rule. Note that this PMSE is constructed from a different number of items in the different sample size combinations. For instance, when $J = 10$ each replication consists of predicting 10 future observables and thus the PMSE is the average of 10 squared difference, while for $J = 300$ each replication consists of predicting 300 future observables and thus the PMSE is the average of 300 squared differences. To be sure, since 100 replications are taken, the average of PMSE over the replications should be fairly reliable and enable the comparison across design conditions for variability in PMSE.

[16]The formation of predictive intervals was also employed where we examined the percent of correct coverage over the replications. However, due to the discrete nature of coverage—in the interval or outside the interval—this proved to be less insightful than the continuous measure of predictive mean square error.

In order to facilitate the analysis of the results in the twelve different parametric design conditions, references will periodically be made to the chart given in table 4. These may be naturally partitioned into three groups: Group One consists of design numbers 1-4 and has the correlation between intercepts and slopes constant at the low value of 0.25; Group Two consists of design numbers 5-8 and has the correlation between intercepts and slopes constant at the medium value of 0.5; and Group Three consists of design numbers 9-12 and has the correlation between intercepts and slopes constant at the high value of 0.75. Within the groups, the level-2 variance varies from 0.25 to 2.0, the intraclass correlation varies between 0.2 and 0.8, and the level-2 covariance increases within the groups as the level-2 variance increases. Table 4 shows that the magnitude of the level-2 covariance is on average higher for the groups with higher correlation between intercepts and slopes.

Results will be tabulated separately for each of the twelve design conditions. Although this creates a plethora of tables, it facilitates the detection of variation across the twelve design conditions. Furthermore, researchers often work with data that is very close to one of these particular design numbers in some aspect, e.g., high intraclass correlation in repeated measures data, and presenting the individual tables will allow such researches to refer to the tables more applicable to their particular interests.

Design number	τ_{00}, τ_{11}	τ_{01}	$r_{u_{0j},u_{1j}}$	ρ
1	0.125	0.03125	0.25000	0.200
2	0.333	0.08330	0.25000	0.400
3	0.75	0.1875	0.25000	0.600
4	2.0	0.50000	0.25000	0.800
5	0.125	0.0625	0.5000	0.200
6	0.333	0.1667	0.5000	0.400
7	0.75	0.3750	0.5000	0.600
8	2.0	1.0000	0.5000	0.800
9	0.125	0.09375	0.75000	0.200
10	0.333	0.25000	0.75000	0.400
11	0.75	0.56250	0.75000	0.600
12	2.0	1.50000	0.75000	0.800

Table 4: Design numbers

3.1 Terrace-Two

The computer code for generating the data was written in XLISP-STAT[17] and the multilevel modeling was done with several altered versions of Terrace-Two.[18] Although many

[17]XLISP-STAT was developed by Luke Tierney and is written in the Xlisp dialect of Lisp, which was developed by David Betz

[18]An XLISP-STAT program written by James Hilden-Minton which incorporates both the EM algorithm and Fisher scoring. As noted by Hilden-Minton, while the latter approach is faster, the EM algorithm exhibits greater stability. Initial estimates are obtained from the first iteration of the EM algorithm, after which point the procedure is switched to Fisher scoring and remains with Fisher scoring until convergence unless Fisher scoring produces estimates outside of the parameter space. See "Terrace-Two User's Guide: An

of the more popular multilevel software packages are faster, the object oriented nature of XLISP-STAT facilitated the amendment and alteration of Terrace-Two in order extend its capability. Defaults such as the maximum number of iterations were changed to allow the number of replications to proceed in the background. Regarding computing time, some of the higher level $J \times n$ sample size combinations were very computer intensive, requiring several hours of computing time on Sun Sparc 10 machines. The limiting factor in the simulations was the actual estimation of the multilevel model, which is a function of J, the number of groups, and not $N = Jn$ the total sample sizes. The data simulations and formation of prediction rules after estimation required very little computing time.

4 Results

Tables 5 - 7 below illustrate initial results in comparing the three prediction rules, where the results have been averaged over the twelve parametric design conditions to facilitate this initial discussion. Later the results will be discussed with respect to variation across each of these twelve parametric design conditions. Note that the cells of the tables represent the average PMSE over 1200 replications for the corresponding prediction rule. [19]

J	n=5	n=10	n=25	n=50	n=100
10	0.4112	0.3146	0.2752	0.2651	0.2561
25	0.3807	0.3073	0.2736	0.2626	0.2561
50	0.3793	0.3053	0.2702	0.2588	0.2554
100	0.3782	0.3036	0.2725	0.2597	0.2548
300	0.3775	0.3061	0.2714	0.2602	0.2551

Table 5: Mean MSE for Multilevel Prediction

J	n=5	n=10	n=25	n=50	n=100
10	1.5995	1.5285	1.5469	1.5362	1.5027
25	1.7384	1.7231	1.7350	1.7355	1.6875
50	1.7984	1.7794	1.8106	1.8268	1.7925
100	1.7973	1.8395	1.7906	1.8141	1.7955
300	1.8484	1.8468	1.7906	1.8436	1.8403

Table 6: Mean MSE for Prior Prediction

The information in Tables 5 - 7 clearly indicates that the multilevel method consistently produces the lowest PMSE across each of the $J \times n$ sample size combinations. Specifically, the multilevel prediction rule produced the lowest average PMSE in 24 of the 25 possible

XLISP-STAT Package for Estimating Multi-Level Models" by Afshartous & Hilden-Minton for a full description of Terrace-Two. Software and manuals accessible via World Wide Web site http://www.stat.ucla.edu.

[19] Twelve parameter design conditions and one hundred replications per design condition, yielding 1200.

J	n=5	n=10	n=25	n=50	n=100
10	0.4581	0.3170	0.2750	0.2649	0.2562
25	0.4378	0.3137	0.2744	0.2628	0.2562
50	0.4496	0.3140	0.2712	0.2590	0.2555
100	0.4512	0.3124	0.2739	0.2600	0.2549
300	0.4500	0.3153	0.2739	0.2604	0.2552

Table 7: Mean MSE for OLS Prediction

$J \times n$ combinations, the only exception being the $J = 10$, $n = 50$ case where the OLS prediction rule produced a nearly identical PMSE to that of the multilevel prediction rule (0.2640 versus 0.2651 respectively). As expected, the differential in PMSE between the multilevel and OLS prediction rules becomes less as the group size n increases, a result of the increased reliability of the OLS prediction in such cases. Note that an increase in the number of groups should have little if any effect on the OLS prediction rule, for this method produces prediction independently in each group. As the group size n increases, however, the OLS prediction rule produces PMSEs very similar to that of the multilevel rule, albeit consistently higher.

The prior prediction rule consistently performs the worst of the three methods, and very much so in absolute terms, more than a full unit higher in PMSE in all $J \times n$ combinations. Although increasing the group size n has little effect on the prior prediction rule, there is a considerable rise in PMSE for the prior prediction rule as the number of groups J rises. Recall that the prior prediction method utilizes the rule $\hat{y}_{*j} = W_j\hat{\gamma}$. Bassiri (1988) demonstrated that an increase in J is beneficial with respect to estimation of γ, while here it seems that the prior prediction rule—the performance of which solely depends on our estimation of γ—performs worse when J is increased. Although this may seem contradictory, it is a manifestation of the dangers of using a grand mean to predict at the individual level. For instance, our estimate of $\hat{\gamma}$ is formed via equation 15 which is a sum over J groups. For small J values, this would be fairly representative of the space of groups, whereas for large J this would be less so since the sum would involve many more terms, each sum with its own values for group specific information such as V_j. Thus, as J increases, the chances of mis-predicting within a particular group increases, leading to the exhibited behavior of the prior prediction rule. These results are further illustrated via several graphical displays employing side-by-side boxplots.

Figures 1 - 5 show the relative performance of the multilevel and OLS prediction rules for particular combinations of group size and observations per group. Figures were plotted on separate scales such that each particular case could be isolated. The advantage of the multilevel prediction rule over the OLS prediction rule is clearly best for low values of n, e.g., $n = 5$ and $n = 10$. In order to highlight the effect of increased group size n on the prediction rules for given levels of the number of groups J, the same figures have been plotted retaining the same scale as n varies from 5 to 100.

Figures 6 - 10 illustrate the improved PMSE as group size n increases for both the multilevel and OLS prediction rules, for all levels of J. In addition, the narrowing of the differential between the multilevel and OLS prediction rules as n increases is also clear for each level of J. The results of the prior prediction rule have intentionally been omitted from

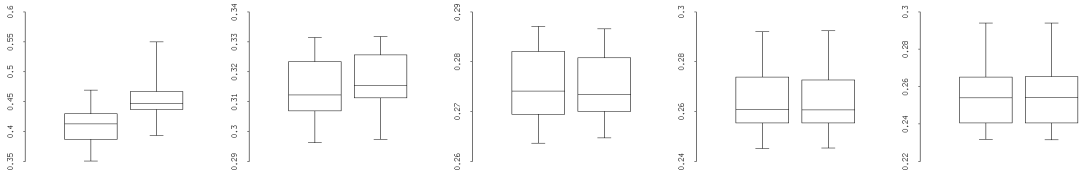

Figure 1: J=10; n=5,10,25,50,100, MSE for ML and OLS

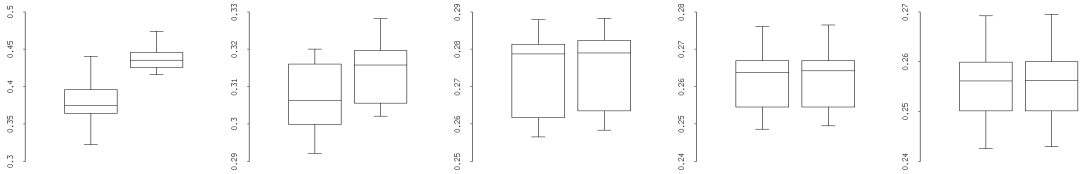

Figure 2: J=25; n=5,10,25,50,100, MSE for ML and OLS

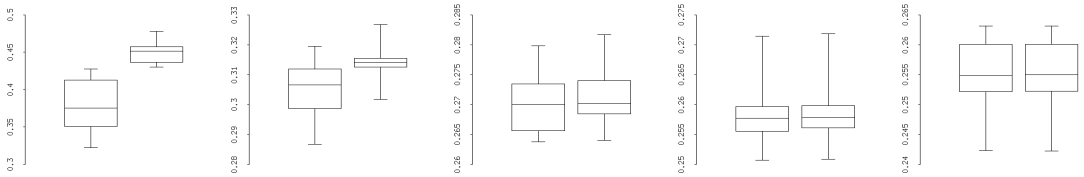

Figure 3: J=50; n=5,10,25,50,100, MSE for ML and OLS

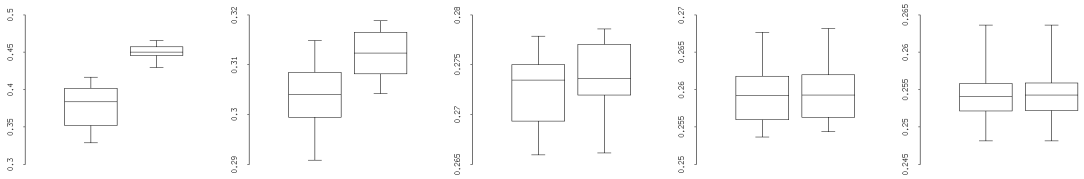

Figure 4: J=100; n=5,10,25,50,100, MSE for ML and OLS

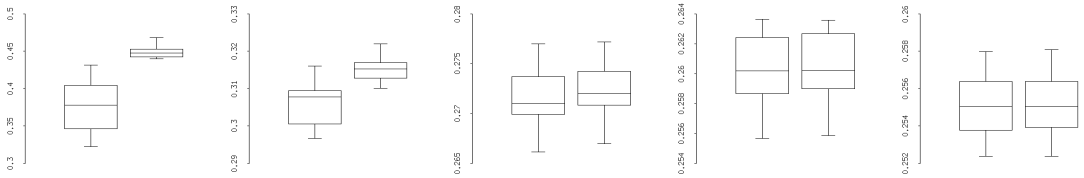

Figure 5: J=300; n=5,10,25,50,100, MSE for ML and OLS

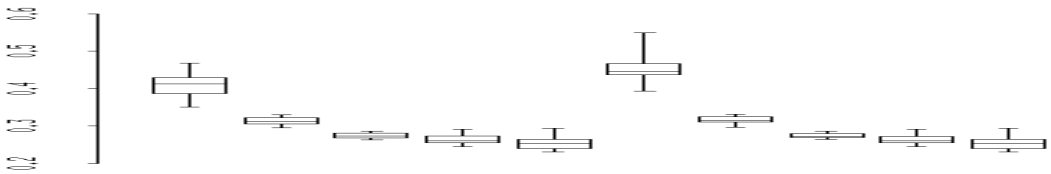

Figure 6: J=10; n=5,10,25,50,100, MSE for ML and OLS

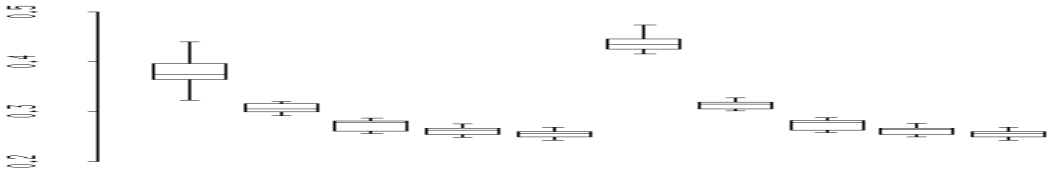

Figure 7: J=25; n=5,10,25,50,100, MSE for ML and OLS

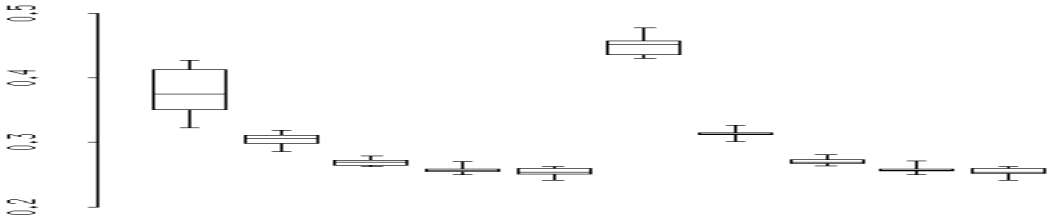

Figure 8: J=50; n=5,10,25,50,100, MSE for ML and OLS

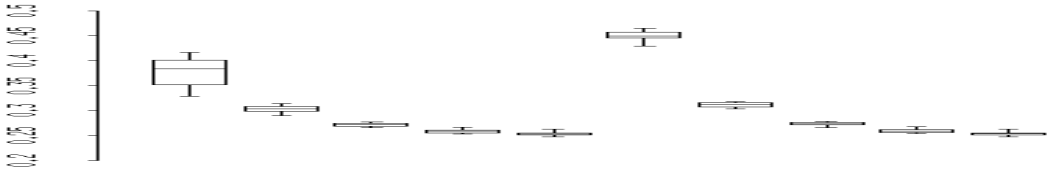

Figure 9: J=100; n=5,10,25,50,100, MSE for ML and OLS

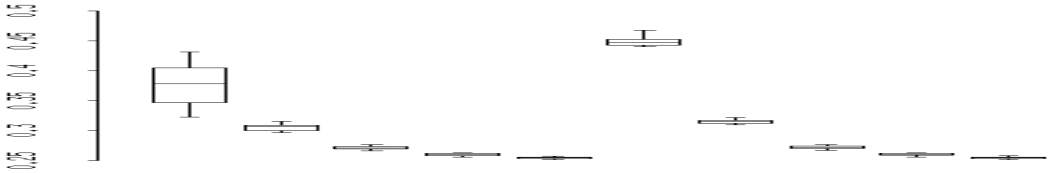

Figure 10: J=300; n=5,10,25,50,100, MSE for ML and OLS

these plots, for the large PMSE values for the prior prediction rules would severely distort the scale and make comparison of the multilevel and prediction rules difficult. Nevertheless, the prior prediction is still of interest. Recall that Table 6 indicated an adverse effect of an increase in J with respect to the PMSE for the prior prediction rule. On the other hand, with respect to the multilevel prediction rule, there is a slight reduction in the overall level of PMSE as J increases. Figures 11 - 15 illustrate this differential effect of increased J for the multilevel and prior prediction methods.[20] Although the reduction in PMSE for the multilevel prediction rule as J increases is slight, the boxplots clearly demonstrate that there is a reduction in the variability of PMSE as J increases for the multilevel prediction rule. For the prior prediction rule, however, not only does the average level of PMSE increase as J increases, the variability in PMSE increases as well.

The use of three-dimensional displays provides additional insight into this disparity between the prediction rules with respect to the effect of J and n. In the three-dimensional plots of Figure 16, both the multilevel and OLS prediction rules produce a PMSE surface that is cleanly sloped downward in the increased n direction. Moreover, there is a slight dip in the J direction as well. The corresponding PMSE surface for the prior prediction rule displays a clearly different relationship, where there is a rather steep slope in the direction of increased J.

These results indicate that a predictive perspective often leads to decisions that differ from those arising from an estimative perspective. Specifically, the results indicate that an increase in group size n is often more beneficial with respect to prediction than an increase in the number of groups J. With respect to the estimation of multilevel model parameters, previous simulation studies (Bassiri 1988, Busing 1994, Mok 1995) indicate that estimation is more improved by increasing the number of groups J instead of the group size n. For example, the sampling distribution of $\hat{\tau}$ is skewed to the right with the true value to the right of the mean of this skewed distribution. Thus, $\hat{\tau}$ will thus be negatively biased in estimating τ. Busing (1994) demonstrated that an increase in the number of groups J was beneficial in reducing this bias, whereas an increase in group size (n) had no affect. With respect to $\hat{\sigma}^2$, Busing obtained relative bias close to zero for all sample design conditions except the smallest sample sizes. This is due to the fact that these level-1 variance estimates are based on the total sample size N, thus this estimate will show little bias as N is large. Finally, focusing on the fixed effects instead of the variance components, Bassiri (1988) demonstrated the improved estimation of γ as the number of groups J increases.

4.1 Parametric variation

The results discussed thus far were obtained by averaging over the twelve parametric design conditions of the simulation design, thereby facilitating this initial comparison of the three prediction rules and the relative effects of J and n on PMSE for the respective prediction rules. Variation over these twelve parametric design conditions is now considered. The individual tables that were averaged to yield Tables 5 - 7 above are presented in Appendices A - C. Figure 17 is a scatterplot matrix of predictive MSE for one particular combination of $J \times n$: $J = 25$ and $n = 5$. Each point represents the average PMSE taken over 100 replications for a given parametric design condition.

[20] The OLS prediction rule is omitted in these figures since, as stated earlier, an increase in J should not effect the OLS prediction rule since prediction for that rule is independent across groups.

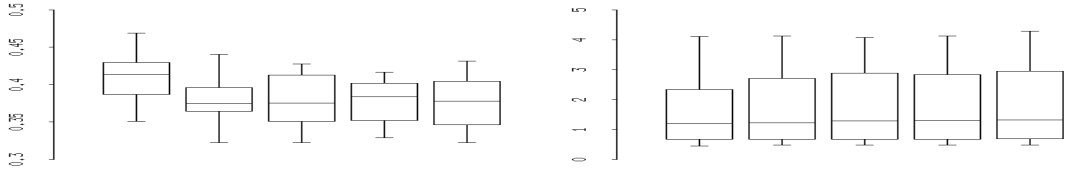

Figure 11: n=5; J=10,25,50,100, 300, MSE for ML and Prior

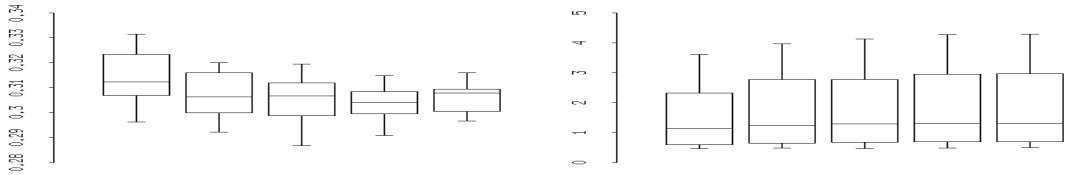

Figure 12: n=10; J=10,25,50,100, 300, MSE for ML and Prior

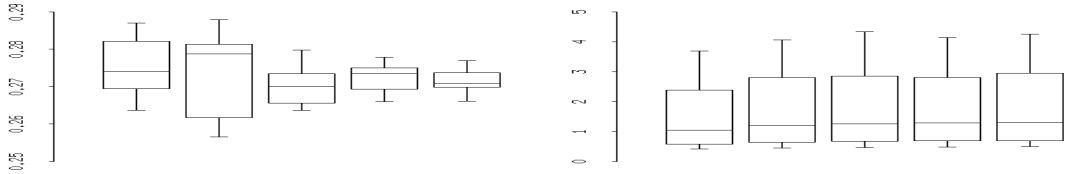

Figure 13: n=25; J=10,25,50,100, 300, MSE for ML and Prior

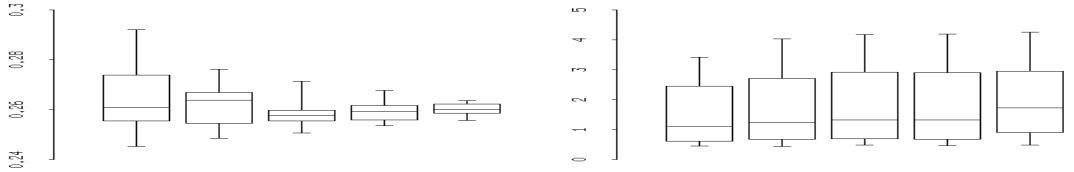

Figure 14: n=50; J=10,25,50,100, 300, MSE for ML and Prior

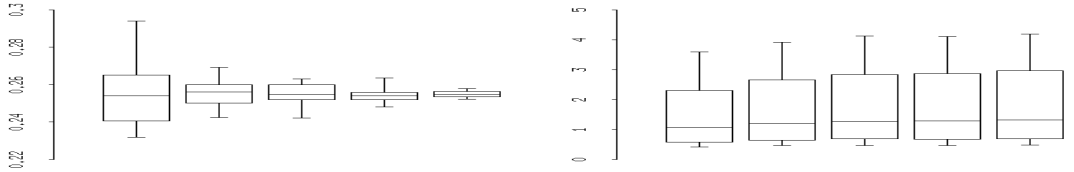

Figure 15: n=100; J=10,25,50,100, 300, MSE for ML and Prior

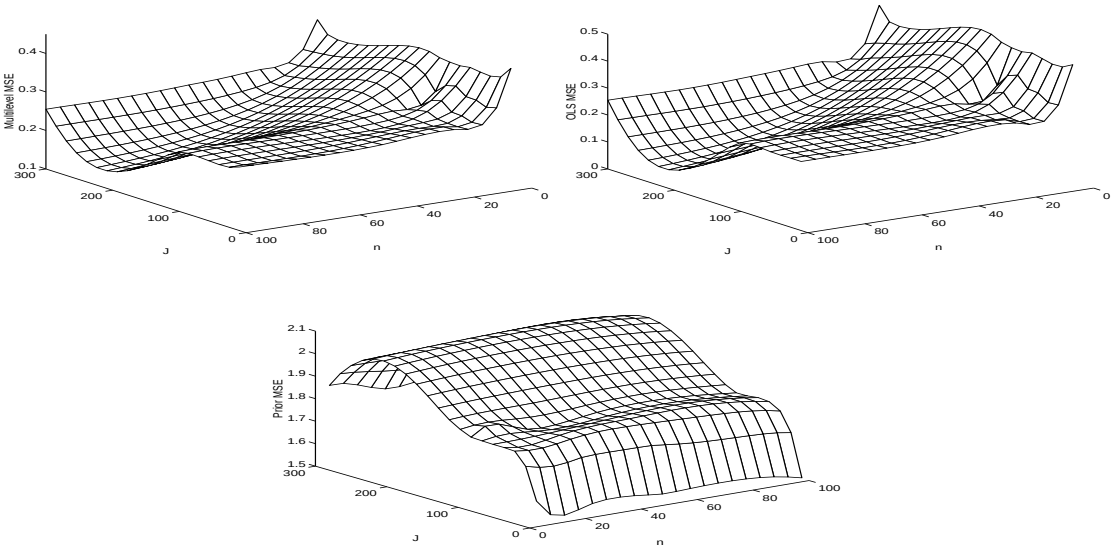

Figure 16: PMSE for Three Prediction Rules

While the multilevel and OLS prediction rules exhibit a relatively even and narrow distribution across the twelve parameter conditions, such is not the case for the prior prediction rule, where the points clearly separate into two distinct groups. Specifically, for three of the twelve parametric design conditions the prior prediction rule performs extremely poorly, with average PMSE over the 100 replications approaching 4.14 whereas the PMSEs for other prediction rules are consistently below 0.5. These three cases are clearly biasing the results that were obtained by averaging over the twelve design conditions, for without them the discrepancy between the prior prediction rule and the other two methods would not be as large. The tables in Appendices A - C indicate that the three parametric conditions in which the prior prediction rule is performing very poorly are conditions numbers 4, 8, and 12 of Table 4. These are the three conditions with a high intraclass correlation coefficient ρ equal to 0.8. In the other nine parametric design conditions, where the intraclass correlation coefficient ranges from 0.2 to 0.6, the prior method is not as far off the multilevel and OLS prediction rules, although its ranking is still a consistent third across all combinations of J and n. Figures 18 -19 illustrate this result for the particular case when $J = 25$ for two particular design conditions. Design #4 represents one of the high intraclass correlation (0.8) situations, while design #1 represents one of the lower intraclass correlation (0.2) situations (See Table 4). With respect to the former, Figure 18 illustrates the very poor performance for the prior prediction rule relative to the other prediction rules, while with respect to the latter Figure 19 illustrates the narrowed differential between the three prediction rules. Similar results hold for other levels of J and other high versus low intraclass correlation design comparisons.

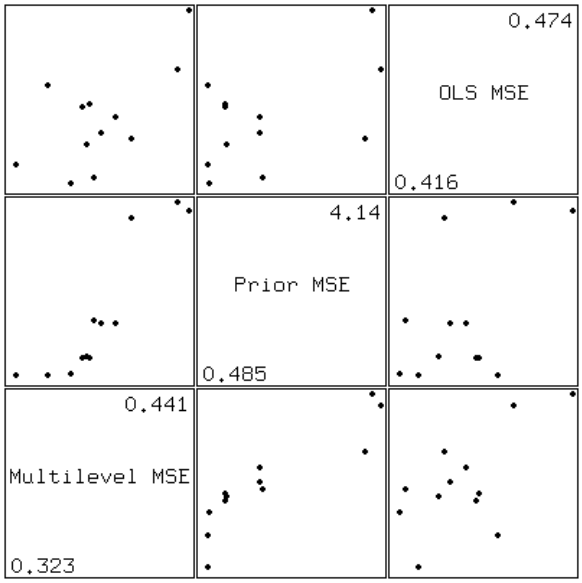

Figure 17: Scatterplot matrix of Predictive MSE; J=25, n=5

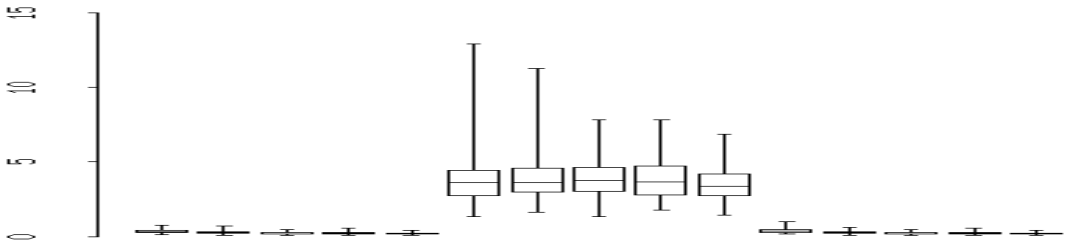

Figure 18: Design #4; J=25; n=5,10,25,50; Multilevel, Prior and OLS PMSE

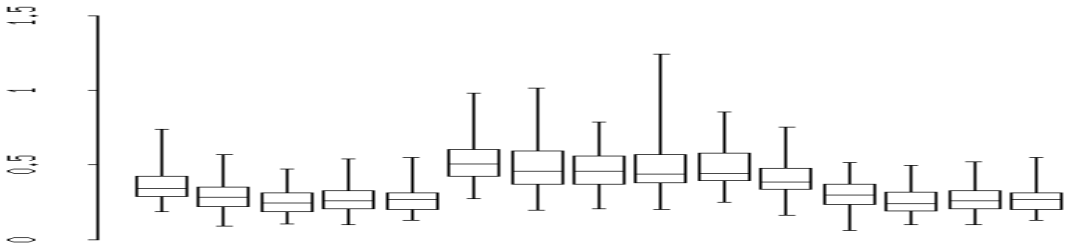

Figure 19: Design #1; J=25; n=5,10,25,50; Multilevel, Prior and OLS PMSE

5 Summary

In summary, we have advocated a predictive approach to multilevel modeling in which the focus lies on the prediction of future observables instead of the characteristics of estimators. Of course, since the prediction rules (OLS, Prior, and Multilevel) are in one-to-one

correspondence to the estimation methods of β_j, these two areas are related. However, one of our main results is that a predictive perspective often leads to decisions that differ from those arising from an estimative perspective. Specifically, the results indicate that an increase in group size n is often more beneficial with respect to prediction than an increase in the number of groups J. With respect to the estimation of multilevel model parameters, previous simulation studies (Bassiri 1988, Busing 1994, Mok 1995) indicate that estimation is more improved by increasing the number of groups instead of the group size.

Three prediction rules were presented and assessed via a Monte Carlo study that extensively covered both the sample size and parameter space. The multilevel prediction rule performed the best across the various design specifications. In addition to the effect of group size mentioned in the previous paragraph, the simulations studies also suggest that the predictive ability of the prior prediction rule is not only poor, but also weakens as the number of groups J increases. These results are further summarized below:

1. The multilevel prediction rule is clearly the best across the $J \times n$ combinations.

2. PMSE is reduced as group size n increases for both the multilevel and OLS prediction rules, for all levels of J

3. The differential in PMSE between the multilevel and OLS prediction rules becomes less as the group size n increases.

4. The prior prediction rule consistently performs the worst of the three prediction rules in absolute terms—more than a full unit higher in PMSE in all $J \times n$ combinations.

5. There is an adverse effect of an increase in J with respect to the PMSE of the prior prediction rule. Not only does the average level of PMSE increase as J increases, the variability in PMSE also increases as well.

6. With respect to the multilevel prediction rule, there is only a slight reduction in the overall level of PMSE as J increases. However, there is a reduction in the variability of PMSE for the multilevel prediction rule as J increases.

7. While the multilevel and OLS prediction rules exhibit a relatively even and narrow distribution across the twelve parameter conditions, such is not the case for the prior prediction rule, where the points clearly separate into two distinct groups. Specifically, for high intraclass correlation (0.8) the prior prediction rule performs extremely poorly. While for lower intraclass correlations, the performance of the prior prediction rule is much closer to that of the multilevel and OLS prediction rules.

A Multilevel Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.3804	0.3210	0.2853	0.2687	0.2396
25	0.3608	0.2958	0.2591	0.2758	0.2693
50	0.3447	0.2963	0.2644	0.2558	0.2519
100	0.3379	0.3004	0.2765	0.2677	0.2544
300	0.3376	0.2968	0.2662	0.2587	0.2558

Table 8: Design #1: Mean MSE for Multilevel Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.4503	0.3304	0.2814	0.2921	0.2643
25	0.3734	0.3051	0.2696	0.2649	0.2547
50	0.3682	0.3196	0.2694	0.2577	0.2565
100	0.3789	0.3051	0.2660	0.2599	0.2549
300	0.3737	0.3089	0.2741	0.2587	0.25714

Table 9: Design #2: Mean MSE for Multilevel Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.4141	0.3062	0.2766	0.2553	0.2519
25	0.3813	0.3157	0.2786	0.2559	0.2507
50	0.4071	0.3112	0.2638	0.2508	0.2534
100	0.3968	0.3010	0.2694	0.2669	0.2611
300	0.3990	0.3094	0.2710	0.26365	0.2568

Table 10: Design #3: Mean MSE for Multilevel Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.4434	0.3145	0.2692	0.2454	0.2418
25	0.4023	0.3166	0.2614	0.2641	0.2588
50	0.4267	0.3126	0.2760	0.2594	0.2525
100	0.4171	0.3149	0.2741	0.2586	0.2482
300	0.4317	0.3161	0.2770	0.2627	0.2554

Table 11: Design #4: Mean MSE for Multilevel Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.3515	0.3057	0.2871	0.2646	0.2319
25	0.3449	0.2923	0.2796	0.2487	0.2585
50	0.3354	0.2972	0.2799	0.2554	0.2632
100	0.3334	0.2909	0.2745	0.2537	0.2538
300	0.3319	0.2989	0.2695	0.2623	0.2535

Table 12: Design #5: Mean MSE for Multilevel Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.4137	0.3077	0.2702	0.2639	0.2625
25	0.3687	0.3192	0.2567	0.2762	0.2612
50	0.3599	0.3016	0.2730	0.2682	0.2534
100	0.3707	0.3096	0.2756	0.2616	0.2637
300	0.3683	0.3025	0.2744	0.2614	0.2535

Table 13: Design #6: Mean MSE for Multilevel Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.3940	0.3085	0.2683	0.2791	0.2665
25	0.3769	0.3077	0.2858	0.2649	0.2692
50	0.3833	0.3127	0.2740	0.2579	0.2424
100	0.4026	0.3030	0.2779	0.2560	0.2521
300	0.3915	0.3096	0.2704	0.2596	0.2544

Table 14: Design #7: Mean MSE for Multilevel Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.4695	0.2963	0.2743	0.2578	0.2395
25	0.4331	0.3201	0.2790	0.2518	0.2578
50	0.4198	0.3005	0.2706	0.2540	0.2622
100	0.4148	0.3074	0.2729	0.2609	0.2557
300	0.4233	0.3131	0.2734	0.2604	0.2524

Table 15: Design #8: Mean MSE for Multilevel Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.3732	0.3176	0.2637	0.2557	0.2567
25	0.3231	0.3018	0.2622	0.2636	0.2499
50	0.3232	0.2868	0.2669	0.2715	0.2511
100	0.3295	0.2929	0.2694	0.2550	0.2523
300	0.3226	0.2984	0.2666	0.2568	0.2541

Table 16: Design #9: Mean MSE for Multilevel Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.4137	0.3103	0.2829	0.2548	0.2942
25	0.3719	0.2981	0.2795	0.2629	0.2427
50	0.3579	0.3061	0.2640	0.2601	0.2587
100	0.3670	0.2987	0.2745	0.2621	0.2514
300	0.3561	0.3032	0.2709	0.2600	0.2580

Table 17: Design #10: Mean MSE for Multilevel Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.4136	0.3259	0.2698	0.2855	0.2751
25	0.3911	0.3130,	0.2880	0.2691	0.2505
50	0.3978	0.3074	0.2715	0.2567	0.2615
100	0.3889	0.3060	0.2669	0.2581	0.2539
300	0.3833	0.3083	0.2710	0.2557	0.2547

Table 18: Design #11: Mean MSE for Multilevel Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.4166	0.3315	0.2740	0.2580	0.2488
25	0.4409	0.3025	0.2832	0.2534	0.2500
50	0.4281	0.3114	0.2692	0.2580	0.2580
100	0.4013	0.3132	0.2728	0.2561	0.2561
300	0.4107	0.3075	0.2719	0.2626	0.2560

Table 19: Design #12: Mean MSE for Multilevel Prediction

B Prior Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.4806	0.4715	0.4487	0.4734	0.4348
25	0.5232	0.4914	0.4691	0.4895	0.4858
50	0.4933	0.4817	0.4785	0.4839	0.4809
100	0.4978	0.5048	0.5073	0.5056	0.4787
300	0.5005	0.5005	0.5025	0.4961	0.4972

Table 20: Design #1: Mean MSE for Prior Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.8710	0.8114	0.7442	0.8394	0.8012
25	0.8623	0.8366	0.8710	0.8983	0.8719
50	0.8828	0.9121	0.9043	0.8799	0.8797
100	0.8989	0.9057	0.8944	0.8749	0.8936
300	0.9221	0.9252	0.9113	0.9086	0.9114

Table 21: Design #2: Mean MSE for Prior Prediction

J	n=5	n=10	n=25	n=50	n=100
10	1.4870	1.4911	1.4736	1.4119	1.3574
25	1.5859	1.5630	1.5545	1.5693	1.5867
50	1.7201	1.6886	1.7247	1.7406	1.6561
100	1.7642	1.6943	1.6815	1.7316	1.6928
300	1.7357	1.6984	1.7463	1.7414	1.7576

Table 22: Design #3: Mean MSE for Prior Prediction

J	n=5	n=10	n=25	n=50	n=100
10	3.1634	3.6134	3.2633	3.3281	3.2444
25	3.8050	3.9693	3.9611	3.9891	3.6414
50	4.0761	3.8674	4.3487	4.1581	4.1333
100	4.1428	4.2827	4.1555	4.0683	4.0672
300	4.2913	4.2120	4.2306	4.1935	4.1962

Table 23: Design #4: Mean MSE for Prior Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.4629	0.4737	0.4770	0.4636	0.4316
25	0.4899	0.4856	0.5019	0.4471	0.4773
50	0.4911	0.4988	0.5047	0.5092	0.5134
100	0.4971	0.4917	0.4965	0.4920	0.4966
300	0.4975	0.5011	0.5002	0.5057	0.4931

Table 24: Design #5: Mean MSE for Prior Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.9439	0.7897	0.8429	0.7624	0.6993
25	0.8629	0.9232	0.8533	0.9375	0.8418
50	0.8908	0.8778	0.9202	0.9373	0.9111
100	0.8882	0.9067	0.8872	0.9210	0.8960
300	0.9198	0.9207	0.9118	0.9063	0.9247

Table 25: Design #6: Mean MSE for Prior Prediction

J	n=5	n=10	n=25	n=50	n=100
10	1.4833	1.4497	1.2822	1.4694	1.3994
25	1.6436	1.6584	1.7077	1.6836	1.7083
50	1.7531	1.7077	1.7473	1.7507	1.7448
100	1.7301	1.7171	1.6959	1.7402	1.7052
300	1.7236	1.7106	1.7468	1.7306	1.7529

Table 26: Design #7: Mean MSE for Prior Prediction

J	n=5	n=10	n=25	n=50	n=100
10	4.1243	3.1577	3.6729	3.4331	3.4315
25	4.1371	3.8842	4.0711	3.7130	3.8260
50	4.0242	4.1084	4.1728	4.18561	3.9340
100	3.9333	4.2321	4.1465	4.1989	4.1038
300	4.1810	4.2972	4.2700	4.2345	4.1889

Table 27: Design #8: Mean MSE for Prior Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.5202	0.4955	0.4281	0.4850	0.4793
25	0.4846	0.4878	0.4540	0.4722	0.4925
50	0.4926	0.4780	0.4900	0.5228	0.4868
100	0.4925	0.4955	0.4916	0.4789	0.4968
300	0.4965	0.5044	0.5008	0.5006	0.4983

Table 28: Design #9: Mean MSE for Prior Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.8436	0.7217	0.7159	0.7494	0.8251
25	0.8829	0.8133	0.8164	0.8789	0.8210
50	0.8864	0.8813	0.8707	0.9109	0.9213
100	0.9086	0.9192	0.9156	0.8981	0.8870
300	0.9124	0.9145	0.9097	0.9109	0.9341

Table 29: Design #10: Mean MSE for Prior Prediction

J	n=5	n=10	n=25	n=50	n=100
10	1.5218	1.4944	1.5176	1.6028	1.3165
25	1.6007	1.6672	1.6476	1.7074	1.5725
50	1.7818	1.7084	1.5967	1.7319	1.7387
100	1.7361	1.7597	1.7018	1.7362	1.705
300	1.7515	1.7468	1.7201	1.7265	1.733

Table 30: Design #11: Mean MSE for Prior Prediction

J	n=5	n=10	n=25	n=50	n=100
10	3.2924	3.3717	3.6959	3.4160	3.6119
25	3.9824	3.8971	3.9128	4.0404	3.9247
50	4.0887	4.1428	3.9687	4.1102	4.1102
100	4.0784	4.1649	3.9137	4.1238	4.1238
300	4.2488	4.2298	4.1651	4.2681	4.1954

Table 31: Design #12: Mean MSE for Prior Prediction

C OLS Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.5507	0.3282	0.2845	0.2673	0.2398
25	0.4162	0.3056	0.2610	0.2766	0.2696
50	0.4782	0.3128	0.2674	0.2567	0.2521
100	0.4452	0.3188	0.2786	0.2682	0.2547
300	0.4423	0.3129	0.2674	0.2591	0.2560

Table 32: Design #1: Mean MSE for OLS Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.5367	0.3249	0.2817	0.2926	0.2645
25	0.4430	0.3156	0.2700	0.2646	0.2548
50	0.4307	0.3269	0.2697	0.2581	0.2566
100	0.4512	0.3124	0.2662	0.2601	0.2551
300	0.4479	0.3168	0.2749	0.2589	0.2571

Table 33: Design #2: Mean MSE for OLS Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.4121	0.3057	0.2754	0.2559	0.2517
25	0.4333	0.3210	0.2792	0.2560	0.2509
50	0.4533	0.3151	0.2641	0.2509	0.2534
100	0.4464	0.3043	0.2697	0.2669	0.2611
300	0.4523	0.3129	0.2711		0.2568

Table 34: Design #3: Mean MSE for OLS Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.4447	0.3123	0.2689	0.2455	0.2420
25	0.4316	0.3163	0.2614	0.2643	0.2589
50	0.4415	0.3138	0.2762	0.2593	0.2525
100	0.4467	0.3156	0.2741	0.2586	0.2482
300	0.4685	0.3178	0.2772	0.2628	0.2554

Table 35: Design #4: Mean MSE for OLS Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.4489	0.3127	0.2867	0.2643	0.2317
25	0.4495	0.3022	0.2790	0.2496	0.2583
50	0.4330	0.3168	0.2817	0.2556	0.2632
100	0.4507	0.3058	0.2776	0.2544	0.2540
300	0.4659	0.3168	0.2719	0.2626	0.2537

Table 36: Design #5: Mean MSE for OLS Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.4727	0.3103	0.2708	0.2635	0.2628
25	0.4423	0.3282	0.2584	0.2758	0.2613
50	0.4342	0.3093	0.2740	0.2683	0.2534
100	0.4529	0.3189	0.2765	0.2617	0.2637
300	0.4405	0.3117	0.2756	0.2614	0.2535

Table 37: Design #6: Mean MSE for OLS Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.3939	0.3150	0.2700	0.2783	0.2666
25	0.4184	0.3108	0.2853	0.2649	0.2692
50	0.4398	0.3160	0.2741	0.2577	0.2423
100	0.4629	0.3073	0.2784	0.2561	0.2520
300	0.4484	0.3145	0.2707	0.2597	0.2544

Table 38: Design #7: Mean MSE for OLS Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.4623	0.2975	0.2735	0.2578	0.2394
25	0.4546	0.3207	0.2792	0.2517	0.2579
50	0.4507	0.3018	0.2705	0.2541	0.2623
100	0.4465	0.3091	0.2731	0.2609	0.2557
300	0.4540	0.3143	0.2736	0.2604	0.2524

Table 39: Design #8: Mean MSE for OLS Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.4418	0.3264	0.2648	0.2543	0.2573
25	0.4229	0.3162	0.2659	0.2644	0.2495
50	0.4526	0.3138	0.2699	0.2719	0.2516
100	0.4504	0.3175	0.2720	0.2565	0.2525
300	0.4410	0.3221	0.2696	0.2574	0.2542

Table 40: Design #9: Mean MSE for OLS Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.4582	0.3233	0.2800	0.2553	0.2943
25	0.4294	0.3058	0.2818	0.2632	0.2430
50	0.4650	0.3148	0.2654	0.2604	0.2589
100	0.4652	0.3110	0.2757	0.2623	0.2515
300	0.4439	0.3173	0.2725	0.2601	0.2581

Table 41: Design #10: Mean MSE for OLS Prediction

J	n=5	n=10	n=25	n=50	n=100
10	0.4333	0.3161	0.2702	0.2854	0.2754
25	0.4385	0.3187	0.2884	0.2691	0.2506
50	0.4539	0.3124	0.2720	0.2571	0.2614
100	0.4659	0.3124	0.2676	0.2582	0.2539
300	0.4477	0.3162	0.2716	0.2559	0.2547

Table 42: Design #11: Mean MSE for OLS Predictio

J	n=5	n=10	n=25	n=50	n=100
10	0.4418	0.3319	0.2735	0.2580	0.2485
25	0.4743	0.3028	0.2830	0.2533	0.2499
50	0.4619	0.3145	0.2696	0.2582	0.2582
100	0.4303	0.3152	0.2732	0.2562	0.2562
300	0.4479	0.3102	0.2722	0.2628	0.2560

Table 43: Design #12: Mean MSE for OLS Prediction

References

Afshartous, David (1997). *Prediction in Multilevel Models*, unpublished Ph.D. dissertation, UCLA.

Afshartous, David. & Hilden-Minton, James (1996). "TERRACE-TWO: An XLISP-STAT Software Package for Estimating Multilevel Models: User's Guide," *U.C.L.A Department of Statistics Technical Report.*

Atchinson J. (1975). "Goodness of Prediction Fit," *Biometrika*, 62, pp.547-554.

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

Butler, Ronald W. (1986). "Predictive Likelihood with Applications," *Journal of the Royal Statistical Society, Series B*, 48, pp.1-38.

Busing, F. (1993). "Distribution Characteristics of Variance Estimates in Two-level Models," Technical Report PRM 93-04, Department of Psychometrics and Research Methodology, University of Leiden, Leiden, Netherlands.

Chipman, J.S. (1964). "On Least Squares with Insufficient Observations," *Journal of the American Statistical Association* , 59, pp.1078-1111.

de Leeuw, Jan & Kreft, Ita., eds. (2002). *Handbook of Multilevel Quantitative Analysis.* Boston, Dordrecht, London: Kluwer Academic Publishers. *In Press.*

de Leeuw, Jan & Kreft, Ita. (1995). "Questioning Multilevel Models," *Journal of the Educational and Behavioral Statistics*, 20, pp.171-189.

de Leeuw, Jan & Kreft, Ita (1986). "Random Coefficient Models for Multilevel Analysis," *Journal of Educational Statistics*, 11, pp.57-86.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society, Series B*, 39, pp.1-8.

Gelfand, A., & Smith, A. (1990). "Sampling Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, pp.398-409.

Geisser, Seymour (1971). "The Inferential Use of Predictive Distributions," in *Foundations of Statistical Inference*, eds., V.P. Godambe and D.A. Sprott, pp.456-469. Toronto: Holt, Rhinehart, and Winston.

Geisser, S. (1979). "A Predictive Approach to Model Selection," *Journal of the American Statistical Association* , 74, pp.153-160.

Goldberger, A.S. (1962). "Best Linear Unbiased Prediction in the General Linear Model," *Journal of the American Statistical Association*, 57, p.369-375.

Goldstein, H. (1986). "Multilevel Mixed Linear Model Analysis using Iterative Generalized Least Squares," *Biometrika*, 78, pp45-51.

Gotway, C. & Cressie, N. (1993). "Improved Multivariate Prediction under a General Linear Model," *Journal of Multivariate Analysis*, 45, 56-72.

Gray, J., Goldstein, H., Thomas, S. (2001). "Predicting the Future: The Role of Past Performance in Determining Trends in Institutional Effectiveness," *Multilevel Models Project web page, www.ioe.ac.uk/multilevel./.*

Harville, David A. (1985). "Decomposition of Prediction Error," *Journal of the American Statistical Association*, 80, p.132-138.

Harville, David A. (1976). "Extension of the Gauss Markov Theorem to Include the Estimation of Random Effects," *Annals of Statistics*, 4, p.384-396.

Hilden-Minton, James (1995). *Multilevel Diagnostics for Mixed and Hierarchical Linear Models*, unpublished Ph.D. dissertation, UCLA.

Hedeker, D., & Gibbons, R. (1996). "MIXOR: A Computer Program for Mixed-effects Ordinal Probit and Logistic Regression Analysis," *Computer Methods and Programs in Biomedicine*, 49, pp.157-176.

Kullback, S. & Leibler, R.A. (1951). "On Information and Sufficiency," *Annals of Mathematical Statistics*, 22, pp.525-540.

Larimore, W.E. (1983). "Predictive Inference, Sufficiency, Entropy and an Asymptotic Likelihood Principle," *Biometrika*, 70, pp.175-182.

Levy, Martin S., & Perng S.K. (1986). "An Optimal Prediction Function for the Normal Linear Model," *Journal of the American Statistical Association*, 81, p.196-198.

Lindley, D.V., & Smith, A.F.M. (1972). "Bayes estimates for the linear model," *Journal of the Royal Statistical Society, Series B*, 34, pp.1-41.

Liski, E.P., & Nummi, T. (1996). "Prediction in Repeated-Measures Models with Engineering Applications," *Technometrics*, 38, 25-36.

Littell, R., Milliken, G., Stroup, W., & Wolfinger, R. (1996). *SAS System for Mixed Models.* Cary, NC: SAS Institute Incorporated.

Longford, N.T. (1988). "Fisher Scoring Algorithm for Variance Component Analysis of Data with Multilevel Structure," in R.D. Bock (ed.), Multilevel Analysis of Educational Data (pp.297-310). Orlando, FL: Academic Press.

Pfefferman, David. (1984). "On Extensions of the Gauss-Markov Theorem to the Case of Stochastic Regression Coefficients," *Journal of the Royal Statistical Society, Series B*, 46, p.139-148.

rao-4 Rao, C.R. (1965b). *Linear Statistical Inference and its Applications, 2nd Edition.* New York: Wiley.

Rao, C.R. (1973). *Linear Statistical Inference and its Applications, 2nd Edition.* New York: Wiley.

Rao, C.R. (1987). "Prediction of Future Observations in Growth Curve Models," *Statistical Science*, 2, 434-471.

Raudenbush, S.W., Bryk, A.S., (2002). *Hierarchical Linear Models, 2nd ed.*, Thousand Oaks: Sage Publications.

Raudenbush, S.W., Bryk, A.S., Cheong, Y., & Congdon, R.T. (2000). *HLM 5: Hierarchical Linear and Nonlinear Modeling.* Chicago: Scientific Software International.

Robinson, G.K. (1991). "That BLUP is a Good Thing," *Statistical Science*, 6, 15-51.

Seltzer, M. (1993). "Sensitivity Analysis for Fixed Effects in the Hierarchical Model: A Gibbs Sampling Approach," *Journal of Educational and Behavioral Statistics*, 18(3), pp.207-235.