A Predictive Density Approach to Predicting a Future Observable 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: A predictive density function g^* is obtained for the multilevel model which is optimal in minimizing a criterion based on Kullback-Leibler divergence for a restricted class of predictive densities, thereby extending results for the normal linear model (Levy & Perng 1986). Based upon this predictive density approach, three prediction methods are examined: Multilevel, Prior, and OLS. The OLS prediction method corresponds to deriving a predictive density separately in each group, while the Prior prediction method corresponds to deriving a predictive density for the entire model. The Multilevel prediction method merely adjusts the Prior prediction method by employing a well known shrinkage estimator from multilevel model estimation. Multilevel data is

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

simulated in order to assess the performance of these three methods. Both predictive intervals and predictive mean square error (PMSE) are used to assess the adequacy of prediction. The multilevel prediction method outperforms the OLS and prior prediction methods, somewhat surprising since the OLS and Prior prediction methods are derived from the Kullback-Leibler divergence criterion. This suggests that the restricted class of of predictive densities suggested by Levy & Perng for the normal linear model may need to be expanded for the multilevel model.

KEY WORDS: prediction, predictive density, multilevel model

1 Introduction

A basic problem in predictive inference involves the prediction of a future observable Z based on the observed data Y in some passed experiment. Moreover, Z need not arise from the same stochastic model as Y. One approach to this problem is to attempt to "estimate" the stochastic model from which Z arises. Given such an estimate, there exist several options for predicting the future observable, e.g., the expected value of the stochastic process. Many authors have investigated this approach, often labeled the predictive density or predictive likelihood method. (Levy & Perng, 1986; Butler, 1986; Geisser, 1971). Another approach is to forgo density estimation and seek to minimize some expected loss function, often within some prescribed class of predictors. (Rao, 1987; Gotway & Cressie, 1993; Goldberger, 1962). Optimal predictors for both approaches have been derived for the general linear model. Moreover,

there exist extensions to the multivariate case (Guttman & Hougaard, 1985; Keyes & Levy, 1996). The purpose of this paper is to extend the optimal predictive density results to the multilevel model. The outline of this paper is as follows: In section 1.1 we review the notation of the multilevel model, in section 2 we present the predictive density approach and the main result by Levy & Perng (1986) for the general linear model. In section 2.1 - 2.3 we develop and apply this result to the multilevel model, thereby obtaining three predictive densities with which to predict a future observation in a hierarchical dataset. In section 2.4 we describe a simulation study to assess the predictive performance of these three densities, in section 3 we present the results, and finally in section 4 we provide a brief summary and directions for future research.

1.1 The Multilevel Model

Multilevel modeling is a statistical technique designed to facilitate inferences from hierarchical data. A given data point y_{ij} represents the ith case in the jth unit, e.g., the ith student in the jth school for educational data. The multilevel model prediction problem—in its simplest form—consists of predicting a future observable y_{*j}, i.e., a future case of the jth group. For a full review of the multilevel model see Bryk & Raudenbush (1992). We shall restrict this discussion to the simple case of primary units grouped within secondary units and periodically refer to the applied example of students (level-1) grouped within schools (level-2). For example, we may have J schools, where the jth school contains n_j students. The basic multilevel model has 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}. In multilevel modeling, some or all of the level-1 coefficients, β_j, are random variables, and 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.

Combining equations (1) and (2) yields the single equation model:

$$Y_j = X_jW_j\gamma + X_ju_j + r_j \tag{3}$$

which may be viewed as a special case of the mixed linear model, with fixed effects γ and random effects u_j.[2] Thus, marginally, y_j has expected value $X_jW_j\gamma$ 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,

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

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.*,

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

Full or restricted maximum likelihood may be applied to this function to produce estimates of σ^2, τ, and γ. These estimates may in turn be employed in various approaches to to produce estimates of the the level-1 coefficients β_j.[3] For a full review of estimation in multilevel models see Raudenbush & Bryk (2002). Although multilevel model estimation 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 we shall employ a predictive density approach to this problem.

2 Predictive Density Approach

Let $f(y;\theta)$ denote the density function for Y and $g(z|y,\theta)$ denote the density function of Z conditioned upon having observed Y. The forms of f and g are assumed known, they are not necessarily the same, and they share the common parameter θ which belongs to some parameter space Θ. Hence the past experiment is informative for the future. A prediction function $s(z;y)$ for z is an estimator of $g(z|y,\theta)$, and if s is a density we call it a predictive density.

[3]The term "estimation" is being used somewhat loosely when speaking of an estimate of β_j since β_j is a random variable. One may consider an estimate of the random variable β_j as an estimate of the mean of its distribution.

Levy & Perng (1986) discuss this problem in the context of the general linear model: Consider an n-dimensional random vector Y and the m-dimensional random vector Z, where $Y = X\beta + \epsilon$ and $Z = W\beta + \tau$, with the usual independence and constant variance variance assumptions for the error terms ($\epsilon \sim n(0, \sigma^2 I_n)$ and $\tau \sim n(\sigma^2 I_m)$). Here $\beta \in \Omega_\beta \subset R^p$ is an unknown $p \times 1$ vector of regression coefficients while σ^2 is an unknown but positive scalar. It is further assumed that ϵ and τ are independent. Letting $p_n(y; X, \beta, \sigma^2)$ and $p_m(z; W, \beta, \sigma^2)$ denote the multivariate normal density functions of Y and Z, respectively, we have

$$\begin{aligned} p_n(y; X, \beta, \sigma^2) &= n(X\beta, \sigma^2 I_n) \\ p_m(z; W, \beta, \sigma^2) &= n(W\beta, \sigma^2 I_m) \end{aligned}$$

Under Kullback-Leibler information loss,[4] Levy & Perng (1986) derive an optimal estimator for the density of Z within a prescribed class of density estimators. Levy & Perng restrict the collection of possible density estimators to a subset of prediction densities, Ψ, and the density within this subset which minimizes the Kullback-Leibler measure is selected. Specifically, they consider the statistics defined by

$$t = t(y, z) = (z - W\hat{\beta})/(n^{1/2}\hat{\sigma}) \tag{6}$$

where $\hat{\beta} = (X'X)^{-1}X'y$ and $\hat{\sigma}^2 = (y - X\hat{\beta})'(y - X\hat{\beta})/n$ are the maximum likelihood estimates of $\hat{\beta}$ and $\hat{\sigma}^2$, respectively. Then Ψ is defined as the collection

[4] Kullback-Leibler information measure was proposed by Atchinson (1975) as a general prediction measure; a discussion of the motivations and properties of this criterion may be found in Larimore (1983).

of all predictive densities that are functions of the statistic t, i.e.,

$$\Psi = \{s(z; W, y, X) : s(z; W, y, X) = g(t(y, z))\}, \tag{7}$$

where g is any probability density function. Two reasons for restricting attention to this class are provided: 1) It contains several commonly used predictors, and 2) the statistic $t(y, z)$ used to define Ψ results from a sequence of data reductions by applying the invariance principle under reasonable groups of transformations. They elaborate by demonstrating maximal invariance with respect to specific groups of transformations; see Levy & Perng (1986, p.197) for further details.

Recall, if $s(z; W, y, X)$ is a predictive density estimate for $p_m(z; W, \beta, \sigma^2)$, then the Kullback-Leibler divergence is defined as:

$$\begin{aligned} D_{\beta,\sigma^2}(p_m, s) &= \int_{R^n} p_n(y; X, \beta, \sigma^2) \int_{R^m} p_m(z; W, \beta, \sigma^2) \\ &\times \; log p_m(z; W, \beta, \sigma^2)/s(z; W, y, X) dz dy \\ &= E_{Y,Z} log[p_m(Z; W, \beta, \sigma^2)/s(Z; W, Y, X)] \end{aligned}$$

Thus, a predictive density s is considered optimal with respect to Kullback-Leibler loss if s minimizes D_{β,σ^2} among all possible predictive densities uniformly with respect to β and σ^2.

Their main result is expressed as follows:

Theorem 1 *Let Ψ, D_{β,σ^2} and t be defined as above. The prediction density*

$$\begin{aligned} g^*(z; W, y, X) &= g^*(t(y, z)) \\ &= st_m(n-p, W\hat{\beta}, n\hat{\sigma}^2 A/(n-p)) \end{aligned}$$

where $A = I_m + W(X'X)^{-1}W'$*, provides the unique minimum of* D_{β,σ^2} *among all* s *in* Ψ *uniformly in* β *and* σ^2*.*

The notation $st_d(k, b, C)$ denotes a multivariate Student-t density function, where d is the dimension, k the degrees of freedom, b the location parameter, and C the dispersion matrix, and density function as follows:

$$\begin{aligned} st_d(k, b, C) &= \Gamma[(k+d)/2]/[\pi^{d/2}\Gamma[k/2] \\ &\times [det(kC)]^{1/2} 1 + (z-b)'(kC)^{-1}(z-b)^{(k+d)/2}]. \end{aligned}$$

As noted by Levy & Perng (1986), assuming a non-informative diffuse prior for (β, σ^2), $f(\beta, \sigma^2) \propto 1/\sigma^2$, the predictive density g^* may be interpreted as a Bayesian predictive density.

Let us examine this predictive density further by using it to create a predictive interval for z in the case where z has dimension one. For large values of n, our predictive interval is centered around the mean of the predictive density, $\hat{z} = W\hat{\beta}$, with margin of error taken as $1.96n\hat{\sigma}^2 A/(n-p)$, in this case a scalar since we have $A = 1 + W(X'X)^{-1}W$. Upon closer examination, however, we see that this interval is very close to the standard exact prediction interval in linear regression. Specifically, the predictive density variance can be written

as follows:

$$\begin{aligned} n\hat{\sigma}^2 A/(n-p) &= \frac{n}{n-p}[\hat{\sigma}^2 + \hat{\sigma}^2 W(X'X)^{-1}W'] \\ &= \frac{n}{n-p}[\hat{\text{var}}(z) + \hat{\text{var}}(W\hat{\beta})] \\ &= \frac{n}{n-p}[\hat{\text{var}}(z) + \hat{\text{var}}(\hat{z})] \end{aligned}$$

Thus, recalling that our usual exact predictive interval in linear regression has margin of error $t_{n-p,\alpha/2}[\hat{\text{var}}(z) + \hat{\text{var}}(\hat{z})]$, the only difference we obtain by using this predictive density to form a prediction interval is the adjustment of the term $\frac{n}{n-p}$ in the expression above. Hence, the resulting interval based on this optimal predictive density would be wider than the exact predictive interval.

We would like to extend this result to the multilevel model. We shall do this in three different ways. First, we extend the theorem above to each of the J groups in the multilevel model as independent OLS regression equations. Thus, in each of the J groups, the prediction problem is identical to the presentation above. This method is referred to as the OLS Prediction Method. Second, we do not ignore the multilevel structure and write the network of J models as one large model and derive the corresponding predictive density for this model. For reasons that will become clear later, this is called the Prior Prediction Method. Finally, we alter the Prior Prediction Method by utilizing a well-known result for the multilevel model to yield the Multilevel Prediction Method.

2.1 OLS Prediction Method

In this case we emphasize that there is no level-2 model, i.e., we do not model the level-1 β_j coefficients as random variables regressed on level-2 vari-

ables. Instead, we simply have J separate regressions:

$$Y_j = X_j\beta_j + r_j, \tag{8}$$

and desire to predict the future observation in the jth group, y_{*j}:

$$y_{*j} = X_{*j}\beta_j + r_{*j}, \tag{9}$$

If y_{*j} were observed, X_{*j} would represent a row of the X_j design matrix and we have $r_{*j} \sim N(0, \sigma^2)$. Thus, we may immediately apply Theorem 1 above to yield the predictive density for y_{*j}:

$$g^*(y_{*j}; X_{*j}, Y_j, X_j) = st_1(n-p, X_{*j}\hat{\beta}_j, n\hat{\sigma_j}^2 A_j/(n-p)) \tag{10}$$

where $A_j = 1 + X_{*j}(X_j{}'X_j)^{-1}X_{*j}{}'$; $\hat{\beta}_j$ and $\hat{\sigma_j}^2$ are the usual OLS estimates for slope and residual variance.

We employ this predictive density to construct a predictive interval by taking its expected value, $X_{*j}\hat{\beta}_j$, as our point predictor for y_{*j} and use its variance to form our margin of error. Formally, we have the following prediction interval:

$$X_{*j}\hat{\beta}_j \pm t_{n-p,.975}\hat{\sigma}_j[nA_j/(n-p)]^{1/2} \tag{11}$$

where $t_{n-p,.975}$ is the .975 critical value for a t distribution with $n-2$ degrees of freedom.

2.2 Prior Prediction Method

In this case we do not ignore the structure of the data; Instead, we adopt the setup of the multilevel model as discussed earlier. However, we first need to do some re-arranging. We shall manipulate the notation in the multilevel model such that it is presented as a special case of the general linear model. 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:

$$Y = X\beta + r \tag{12}$$

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

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

where Ψ_j is usually $\sigma^2 I_{n_j}$. We may also write the level-2 equation in no-subscript form through similar stacking manipulations:

$$\beta = W\gamma + u \tag{13}$$

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:

$$Y = XW\gamma + Xu + r \tag{14}$$

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

Now, consider a future observable y_{*j}, i.e., a future observation in the jth unit. As before, let the level-1 data corresponding to this observation may be denoted as X_{*j}, a $1 \times p$ row vector. To make the analogy with Levy & Perng (1986) explicit, recall that for the general linear model we had the following

distributions:

$$\begin{aligned} p_n(y; X, \beta, \sigma^2) &= n(X\beta, \sigma^2 I_n) \\ p_m(z; W, \beta, \sigma^2) &= n(W\beta, \sigma^2 I_m) \end{aligned}$$

Similarly, for the multilevel model in stacked form, we now have:

$$\begin{aligned} p_N(y; X, \gamma, \sigma^2) &= n(XW\gamma, XTX' + \Psi) \\ p_1(y_{*j}; X_{*j}, \gamma, \sigma^2) &= n(X_{*j}W_j\gamma, V_*) \end{aligned} \tag{15}$$

where $V_* = X_{*j}\tau {X_{*j}}' + \sigma^2$ and $N = \sum_{j=1}^{J} n_j$ equals the total number of cases (students) across all the units or groups (schools). The corresponding parameter estimate for the multilevel model is now γ instead of β and may be estimated as follows:

$$\begin{aligned} \hat{\gamma} &= (\sum_{j=1}^{J} {W_j}' {X_j}' \hat{V}_j^{-1} X_j W_j)^{-1} \sum_{j=1}^{J} {W_j}' {X_j}' \hat{V}_j^{-1} y_j \\ \hat{V}_j &= \hat{\text{var}}(y_j) = X_j \hat{\tau} {X_j}' + \hat{\sigma}^2 I \end{aligned}$$

where $\hat{\tau}$ and $\hat{\sigma}^2$ must be estimated iteratively via full or restricted maximum likelihood. [5] The estimate above for the fixed effects γ may be interpreted as a generalized linear model (GLM) estimator.[6]. If the dispersion matrix in the multilevel model was diagonal as in the of the normal linear model, we

[5] Procedures such as Fisher Scoring (Longford, 1987), iteratively reweighted generalized least squares (Goldstein, 1986), or the EM algorithm (Dempster, Laird, & Rubin, 1977) 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 (Gelfand et al., 1990; Seltzer, 1993).

[6] de Leeuw & Kreft (1995) discuss how alternative estimates of the fixed effects may be obtained 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.

could directly apply the result of Levy & Perng to obtain the corresponding predictive density for the multilevel model. However, as can readily be seen from equation 15 above, the dispersion matrix has a complicated structure. This problem may be solved, however, by making use of some transformations. In order to simplify presentation, we shall first extend Levy & Perng's assumptions in the case of the normal linear model, and then directly apply this result to the multilevel model. Formally, let us generalize Levy & Perng's case to that of the following:

$$\begin{aligned} p_n(y; X, \beta, \sigma^2) &= n(X\beta, \sigma^2\Sigma) \\ p_m(z; W, \beta, \sigma^2) &= n(W\beta, \sigma^2\Delta) \end{aligned}$$

Assume that Σ and Δ are known matrices of rank n and m, respectively. Let G be an $n \times n$ matrix of rank n such that $\Sigma = G'G$. Similarly, let H be an $m \times m$ matrix of rank m such that $\Delta = H'H$. Let $\tilde{y} = G'^{-1}y$ and let $\tilde{z} = H'^{-1}z$. Similarly, let $\tilde{X} = G'^{-1}X$ and let $\tilde{W} = H'^{-1}W$. Thus, the models above have now been transformed as follows:

$$\begin{aligned} p_n(\tilde{y}; \tilde{X}, \beta, \sigma^2) &= n(\tilde{X}\beta, \sigma^2 I_n) \\ p_m(\tilde{z}; \tilde{W}, \beta, \sigma^2) &= n(\tilde{W}\beta, \sigma^2 I_m) \end{aligned}$$

Hence, we are back in the original format of Levy & Perng's problem and may apply the main result to the transformed variables $\tilde{y}$ and $\tilde{z}$ in order to produce the optimal predictive density for $\tilde{z}$. Note that our corresponding maximum likelihood estimates in the transformed model are now $\hat{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y}$ and $\hat{\sigma}^2 = (\tilde{y} - \tilde{X}\hat{\beta})'(\tilde{y} - \tilde{X}\hat{\beta})/n$. However, it can be readily shown that $\hat{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y$ (Graybill, 1976, p.207).

Following along similar lines as Levy & Perng's original development, define $\tilde{t} = (\tilde{z} - \tilde{W}\hat{\beta})/(n^{1/2}\hat{\sigma})$ and restrict the set of candidate optimal predictive densities for $\tilde{z}$ to $\Psi = \{s(\tilde{z}; \tilde{W}, \tilde{y}, \tilde{X}) : s(\tilde{z}; \tilde{W}, \tilde{y}, \tilde{X}) = g(\tilde{t}(\tilde{y}, \tilde{z}))\}$. Thus, applying Theorem 1, we have the optimal predictive density g^* for $\tilde{z}$ over the restricted set Ψ as follows:

$$\begin{aligned} g^*(\tilde{z}; \tilde{W}, \tilde{y}, \tilde{X}) &= g^*(\tilde{t}(\tilde{y}, \tilde{z})) \\ &= st_m(n-p, \tilde{W}\hat{\beta}, n\hat{\sigma}^2 A/(n-p)) \end{aligned}$$

where $A = I_m + \tilde{W}(\tilde{X}'\tilde{X})^{-1}\tilde{W}'$, provides the unique minimum of D_{β,σ^2} among all s in Ψ uniformly in β and σ^2.

Of course, we are interested in the optimal predictive density of z, not $\tilde{z}$, so we must transform back to the original units. Since we have $\tilde{z} = H'^{-1}z$, this implies that $z = H'\tilde{z}$. Thus, our predictive density for z is as follows:

$$\begin{aligned} g^*(z; W, y, X) &= g^*(t(y, z)) \\ &= st_m(n-p, W\hat{\beta}, n\hat{\sigma}^2 \Delta A/(n-p)) \end{aligned} \tag{16}$$

where once again we emphasize that the estimates of $\hat{\beta}$ and $\hat{\sigma}^2$ above are not the same as that in the original development. Recall that previously we showed that the dispersion term derived by Levy & Perng for the normal linear model may be written as an adjusted exact prediction interval, where the adjustment factor was $\frac{n}{n-p}$. Now, with a little bit of algebra, we demonstrate a similar

result for our more general case:

$$\begin{aligned}
\frac{n\hat{\sigma}^2\Delta A}{n-p} &= \frac{n\hat{\sigma}^2\Delta}{n-p}[I + \tilde{W}(\tilde{X}'\tilde{X})^{-1}\tilde{W}'] \\
&= \frac{n}{n-p}[\hat{\sigma}^2\Delta + \hat{\sigma}^2\Delta\tilde{W}(\tilde{X}'\tilde{X})^{-1}\tilde{W}'] \\
&= \frac{n}{n-p}[\tilde{\text{var}}(z) + \Delta\tilde{\text{var}}(\hat{\tilde{z}})] \\
&= \frac{n}{n-p}[\tilde{\text{var}}(z) + \tilde{\text{var}}(H'^{-1}\hat{z})] \\
&= \frac{n}{n-p}[\tilde{\text{var}}(z) + \Delta^{-1}\tilde{\text{var}}(\hat{z})] \\
&= \frac{n}{n-p}[\tilde{\text{var}}(z) + \tilde{\text{var}}(\hat{z})]
\end{aligned} \tag{17}$$

Thus, as in the general case, this illustrates the dispersion of the predictive density with respect to an adjustment to the margin of error in an exact prediction interval. Let us apply this result to our multilevel model such that we may produce a predictive density for y_{*j}. The only difference in assumptions is that certain components (σ^2 and τ) of the dispersion matrices ($XTX' + \Psi$ and V_*) for y and y_{*j} are unknown and must be estimated. Directly applying the results from equations 16 and 17 above yields the following predictive density:

$$t(y_{*j}; X_{*j}, y, X, W) = st_1(N - q, X_{*j}W_j\hat{\gamma}, \frac{N}{N-q}B_j) \tag{18}$$

where q is the length of γ and $B_j = \hat{var}(y_{*j}) + \hat{var}(\hat{y}_{*j})$ and may be written as:

$$\begin{aligned}
B_j &= X_{*j}\hat{\tau}X_{*j}' + \hat{\sigma_j}^2 + \hat{\text{var}}(X_{*j}W_j\hat{\gamma}) \\
&= X_{*j}\hat{\tau}X_{*j}' + \hat{\sigma_j}^2 + X_{*j}W_j\hat{\text{var}}(\hat{\gamma})W_j'X_{*j}' \\
&= \hat{V}_* + \hat{\sigma_j}^2 + X_{*j}W_j(\sum_{j=1}^{J} W_j'X_jV_j^{-1}X_jW_j)^{-1}W_j'X_{*j}'
\end{aligned}$$

Comparing the main result here with that from the previous section, the center of the prediction density is now $X_{*j}W_j\hat{\gamma}$ instead of $X_{*j}\hat{\beta}_j$. One may view $\hat{\gamma}$ as analogous to $\hat{\beta}_j$ with respect to the application of the theorem, noting that maximum likelihood is satisfied via GLM in the former and OLS in the latter. As before, we employ this predictive density by taking its expected value, $X_{*j}W_j\hat{\gamma}$, as our point predictor for y_{*j} and use the variance to form our margin of error. Formally, we have the following prediction interval:

$$X_{*j}W_j\hat{\gamma} \pm t_{N-q,.975}[NB_j/(N-q)]^{1/2} \tag{19}$$

Readers familiar with multilevel models will recognize that this corresponds to employing the prior estimate of β_j, $\hat{\beta}_j^{Prior} = W_j\hat{\gamma}$, in forming $\hat{y}_{*j} = X_{*j}\hat{\beta}_j^{Prior}$; Hence the term Prior Prediction Method. Similarly, in the previous section we employed the OLS estimate for β_j and obtained the OLS Prediction Method. Although the predictive density above corresponding to the Prior Prediction Method is the optimal predictive density in the sense of Levy & Perng (1986), it behooves the researcher to investigate the effect of using the popular multilevel estimate of β_j in place of either the OLS or prior estimate.

2.3 Multilevel Prediction Method

One of the main results in the multilevel model literature is the shrinkage estimator for β_j, which may be expressed as a weighted combination of the OLS and prior estimate. Intuitively, the higher the reliability of the OLS estimate the the larger the weight attached to the OLS estimate, and vice

versa. Formally, the multilevel model estimate $\hat{\beta_j}^*$ may be written as follows:

$$\hat{\beta_j}^* = \delta_j \hat{\beta}_j + (I - \delta_j) W_j \hat{\gamma} \tag{20}$$

where

$$\delta_j = \hat{\tau}[\hat{\tau} + \hat{\sigma^2}(X_j{}' X_j)^{-1}]^{-1} \tag{21}$$

is the ratio of the parameter variance for β_j (τ) relative to the variance of the OLS estimator for β_j ($\sigma^2(X_j{}' X_j)^{-1}$) 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. See Bryk & Raudenbush (2002) for further details on these shrinkage concepts. Once again, the variance components τ and σ^2 must be estimated iteratively and γ is estimated via the GLS equation of the previous section. The shrinkage estimator above for β_j above which employs the estimator of equation (X) yields the minimum mean square linear unbiased estimator (MMSLUE) of β_j (Harville 1976).[7]

One may also write the multilevel estimate as $\hat{\beta_j}^* = W_j\hat{\gamma} + \hat{u_j}$, where we recall that u_j may be interpreted in the mixed model sense as the random effect of the jth group. With respect to the prediction of y_{*j}, we will now take our predicted value of y_{*j} to be $X_{*j}\hat{\beta_j}^*$, which may also be written as $\hat{y}_{*j} = X_{*j}W_j\hat{\gamma} + X_{*j}\hat{u_j}$. Taking this one step further, we note that Harville (1976) showed that this may also be written as follows:

$$\hat{y}_{*j} = X_{*j}W_j\hat{\gamma} + \hat{V}_{*j}\hat{V}_j^{-1}(y_j - X_jW_j\hat{\gamma}) \tag{22}$$

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

where $\hat{V}_{*j} = \text{cov}(y_{*j}, y_j) = X_{*j}\hat{\tau}X_j{}' + \hat{\sigma}^2$. We note this last representation since it illustrates our prediction as the conditional expectation of y_{*j} given the data Y. Furthermore, Rao (1973, p.522) showed that $\hat{y}_{*j}$ is the best predictor of y_{*j} with respect to the minimum mean square error criterion.

Getting back to our predictive density, we now simply center this predictive density around $X_{*j}\hat{\beta}_j^*$ and form its dispersion in a manner analogous to the previous sections, yielding:

$$t(y_{*j}; X_{*j}, y, X, W) = st_1(N-q, X_{*j}\hat{\beta}_j^*, \frac{N}{N-q}C_j) \tag{23}$$

where q is the length of γ and $C_j = \hat{\text{var}}(y_{*j}) + \hat{\text{var}}(\hat{y}_{*j})$ and may be written as:

$$C_j \ = \Omega_{*j} + M_j(\text{var}(\hat{\gamma}))M_j{}'$$

where $\Omega_{*j} = V_* + V_{*j}V_j{}^{-1}V_{j*}$ and $M_j = X_{*j}W_j - V_{*j}V_j{}^{-1}X_jW_j$.[8] As before, we will form our predictive interval for y_{*j} by centering it around the distribution's mean and using its variance to form the margin of error. Formally, we have the following prediction interval:

$$X_{*j}\hat{\beta}_j^* \pm t_{N-q,.975}[NC_j/(N-q)]^{1/2} \tag{24}$$

We investigate the difference between the OLS, Prior, and Multilevel Prediction methods mentioned above through a simulation study. The design of the simulation study is explained in the next section.

[8] This expression is derived in Liski & Nummi, 1996.

2.4 Simulation Study

Multilevel data is simulated under a variety of design conditions, closely following the simulation study of Busing (1993). As a simplification, we consider a simple 2-level multilevel model with one explanatory variable at each level and equal numbers of units in each group. A two-stage simulation scheme is employed. At the first stage the level-1 random components are generated according to the following equations[9]:

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

The γ's are the fixed effects and are set to a predetermined value; We set them all equal to one as in Busing (1993). The scalar 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. We set the two diagonal elements of τ, τ_{00} and τ_{11}, to be .125 and the off-diagonal covariance term τ_{01} at .03, following one of Busing's major design conditions. This yields in intraclass correlation ρ of 0.2.[10]

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

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \epsilon_{ij} \tag{25}$$

[9] There is a slight abuse of notation here. Previously W_j represented a matrix while here it represents a scalar.

[10] The intraclass correlation is defined as follows: $\rho = \frac{\tau_{00}}{\tau_{00} + \sigma^2}$ and thus measures the degree to which units within the same unit are related.

Table 1: Simulation Specification

$\sigma^2 = .5$
$\tau_{00} = \tau_{11} = .125$
$\tau_{01} = .03$
$X_{ij} \sim N(0,1)$
$W_j \sim N(0,1)$
$J = (25, 50, 100)$
$n = (10, 25, 50)$

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. In summary, our parameter specification for simulating multilevel data are as follows:

The multilevel data is simulated under a variety of specifications for the number of groups (J) and number of units per group (n). Once again following Busing (1993), the number of groups studied are 25, 50, and 100, while the of units per group are 10, 25, and 50. Moreover, one additional "future" observation is generated for each of the J groups. Thus, for the J=100, n=25 design specification, 100 additional observation are generated and set aside. These are the observations that will be predicted; They are not used for estimative purposes.

2.5 Prediction Results

The adequacy of prediction was checked in two ways: predictive intervals and predictive mean square error (PMSE). The predictive interval method is performed as follows. For each of the future observations to be predicted, a predictive interval is formed from the respective predictive distribution and we check whether or not the observation lies in this interval. Thus, for $J = 50$

Table 2: Mean Fractional coverage for Multilevel, Prior, and OLS Prediction Intervals

J	n=10	n=25	n=50
25	.949, .946, .984	.961, .951, .973	.952, .950, .952
50	.956, .952, .989	.956, .953, .96	.951, .951, 955
100	.960, .948, .987	.953, .953, .966	.956, .947, .959

we will have a possible range of 0 to 50 correct predictive intervals. Moreover, to check the variability of such coverage, each of the nine $J \times n$ design conditions are simulated 100 times, each time checking the percent of correct intervals. The data simulations were performed in XLISP-STAT while the multilevel model estimation was performed with TERRACE-TWO.[11] Regarding computing time, it took 14.1 minutes to simulate one hundred J=25, n=10 data sets, estimate the models, and form the desired predictions. The corresponding time for the one hundred J=100, n=50 data sets was one hour and 46 minutes; all computations were performed on a SUN Sparc 10 workstation.

The predictive interval results are given in Table 2.5 below, where we give the mean of the fraction of correct intervals over 100 simulations for each design specification. For instance, the entries in the top left cell shows that for the J=25, n=10 design condition the mean fractional coverage for the multilevel, prior, and OLS predictive intervals were .949, .946, and .984, respectively, over the 100 simulations.

For all three methods, the coverage rate is close to that expected from a theoretical 95% prediction interval. Moreover, there isn't much difference between simple OLS and the multilevel intervals. One, this could be a result

[11] An XLISP-STAT program written by James Hilden-Minton, which incorporates both the EM algorithm and Fisher scoring for parameter estimation. See "Terrace-Two User's Guide: An XLISP-STAT Package for Estimating Multi-Level Models" by Afshartous & Hilden-Minton for a full description of Terrace-Two. XLISP-STAT was developed by Luke Tierney and is written in the Xlisp dialect of Lisp, which was developed by David Betz.

of the wideness of our margin of error, and two, it could be a result of the discreteness of the assessment approach we have employed. In order to get around this problem, we also examine the popular predictive mean square error (PMSE) approach to assessing predictive performance.

For the predictive mean square error (PMSE) approach, we employ the standard technique of taking the average of the sum the squared errors (SSE) of the observed and predicted values. The predicted values are taken as the expected value of our predictive density, varying according to our estimate of β_j. Once again, for each of the nine $J \times n$ design conditions, we calculate our result 100 times. The results are summarized in Table 3, where each entry is the average of PMSE over 100 simulations. The multilevel method is clearly the best, closely followed by the OLS method. As expected, the discrepancy between the multilevel and OLS method becomes less as n increases. Increasing J should have no effect on the OLS method since this method forms predictions separately for each group. On the other hand, an increase in J should decrease PMSE for the multilevel method since the multilevel method uses all of the data; However, this is not entirely confirmed in these simulations, possibly because the increase in J is not large enough to make a difference. Somewhat of a surprise, the prior prediction method performs the worst of the three methods, increasingly worse as J increases. Although the multilevel prediction rule is superior, the differential gain is not incredibly large and does not increase dramatically as the design tends towards smaller J and n, i.e., specifications where we would expect the multilevel prediction rule to further outperform the other methods.

These results are more clearly illustrated in Figure 1-3, where we display boxplots of the distribution of PMSE for the three methods over the 100

Table 3: Mean MSE over 100 Simulations for Multilevel, Prior, and OLS Prediction

J	n=10	n=25	n=50
25	0.2958, 0.4914, 0.3056,	0.2591, 0.4691, 0.2610	0.2758, 0.489, .27666
50	0.2963, 0.4817, 0.3128	0.2644, 0.4785, 0.2674	0.2558, 0.4839, 0.2567
100	0.3005, 0.5048, 0.3188	0.2765, 0.5073, 0.2786	0.2677, 0.5056, 0.2682

simulations. Consider Figure 1 where $J = 25$ is fixed. Starting from the left, the first three boxplots correspond to the PMSE for the Multilevel Prediction Method for $n = 10, n = 25$, and $n = 50$, respectively. The next three boxplots correspond to the PMSE for the Prior Prediction Method for $n = 10, n = 25$, and $n = 50$, respectively. And, finally, the last three boxplots correspond to the PMSE for the OLS Prediction Method for $n = 10, n = 25$, and $n = 50$, respectively. Figures 2 and 3 are arrayed similarly for $J = 50$ and $J = 100$, respectively. The effect of group size n is clear as PMSE decreases within each prediction method as n increases. An exception, however, occurs in Figure 1 for the Multilevel Method. And, once again, the poor performance of the Prior Prediction Method is apparent as the corresponding boxplots have higher medians in all design specifications. Moreover, as indicated by the boxplots, examining the standard-deviation of PMSE over the 100 simulations confirms that the Multilevel Predictive approach is also the least variable.[12]

These results demonstrate that in spite of the fact that the OLS and prior prediction rules are based on predictive densities which are optimal in the sense of the Kullback-Leibler divergence criterion as employed by Levy & Perng (1986), the predictive performance of the multilevel prediction rule is superior. Part of the reason for this result may arise from the fact that we have

[12]For the J=100, n=10 specification, the standard deviations of SSE for the multilevel, prior, and OLS methods are 3.678, 5.357, and 3.982, respectively.

restricted the collection of possible density estimators to a subset of prediction densities. This restriction, although possibly useful for theoretical purposes of density estimation, has clearly failed to produce the best density with respect to predicting future observations. Indeed, Levy & Perng (1986) employ this particular restriction in order to demonstrate their result with respect to several other commonly used predictive densities that also belong to this restricted set of predictive densities; whether this set is a reasonably large collection of predictive densities is not their main concern. For the multilevel model at least, our results indicate that this collection needs to be larger.

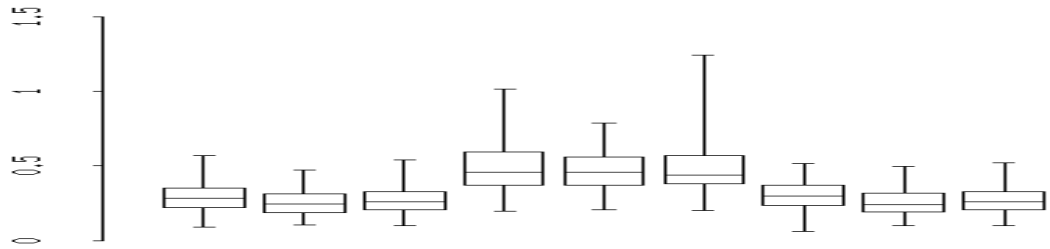

Figure 1: J=25; n=10,25,50 for Multilevel, Prior and OLS MSE over 100 Simulations

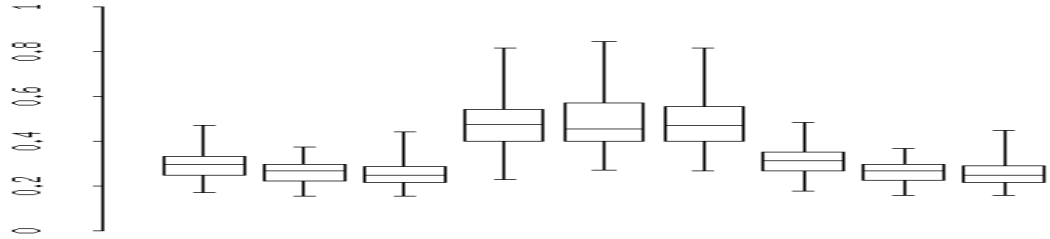

Figure 2: J=50; n=10,25,50 for Multilevel, Prior and OLS MSE over 100 Simulations

3 Summary

A predictive density for the multilevel model has been derived in order facilitate the prediction of future observables in multilevel data. Based upon this predictive density, three prediction methods have been examined: multilevel, prior, and OLS prediction. The OLS prediction method corresponds to deriving a predictive density separately in each group, while the prior prediction method corresponds to deriving a predictive density for the entire model. The multilevel prediction method merely adjusts the Prior prediction method by using a well known result from multilevel model estimation. The adequacy of

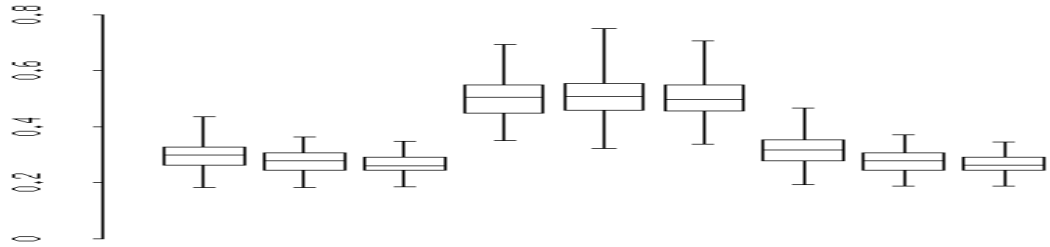

Figure 3: J=100; n=10,25,50 for Multilevel, Prior and OLS MSE over 100 Simulations

prediction has been assessed through both predictive intervals and predictive mean square error (PMSE). Based on simulated multilevel data, the multilevel method is superior. This indicates that for the multilevel model the restriction used by Levy & Perng (1986) in the context of the normal linear model is possibly overly conservative.

The differential gain in prediction for the multilevel method, however, is not incredibly large, nor does this differential gain increase appreciably as the design conditions tend towards smaller J and n, i.e., specifications where we would expect the multilevel method to outperform the OLS method. To be sure, our results might vary if we widen the $J \times n$ space or change other design parameters aside from J and n, e.g., the various parameters of Table 1. In the sequel we explore this enhanced design space and also present a decomposition of prediction error to assess the relative costs of missing data and parameter estimation.

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.

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

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.

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

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

Hilden-Minton, James (1994). *TERRACE-TWO: A New Xlisp-Stat Package for Multilevel Modeling with Diagnostics*, UCLA Statistics Series: (www.stat.ucla.edu)

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.

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

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.

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

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

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

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

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