Behaviormetrika
Vol.31, No.1, 2004, 43–66

# AN APPLICATION OF MULTILEVEL MODEL PREDICTION TO NELS:88

David Afshartous* and Jan de Leeuw**

Multilevel modeling is often used in the social sciences for analyzing data that has a hierarchical structure, e.g., students nested within schools. In an earlier study, we investigated the performance of various prediction rules for predicting a future observable within a hierarchical data set (Afshartous & de Leeuw, 2004). We apply the multilevel prediction approach to the NELS:88 educational data in order to assess the predictive performance on a real data set; four candidate models are considered and predictions are evaluated via both cross-validation and bootstrapping methods. The goal is to develop model selection criteria that assess the predictive ability of candidate multilevel models. We also introduce two plots that 1) aid in visualizing the amount to which the multilevel model predictions are "shrunk" or translated from the OLS predictions, and 2) help identify if certain groups exist for which the predictions are particularly good or bad.

## 1. Introduction

Multilevel modeling is often used in the social sciences for analyzing data that has a hierarchical structure, e.g., students nested within schools (Bryk & Raudenbush, 1992; de Leeuw & Kreft, 2002; Leyland & Goldstein, 2001). Specifically, within each group we have the following level-1 model equation:

$$\mathbf{Y}_j = \mathbf{X}_j\boldsymbol{\beta}_j + \mathbf{r}_j \tag{1.1}$$

Each $\mathbf{X}_j$ has dimensions $n_j \times p$, and $\mathbf{r}_j \sim N(0, \sigma^2\boldsymbol{\Psi}_j)$, with $\boldsymbol{\Psi}_j$ usually taken as $\mathbf{I}_{n_j}$. In multilevel modeling, some or all of the level-1 coefficients, $\boldsymbol{\beta}_j$, are viewed as random variables.[1)] They may also be functions of level-2 (school) variables:

$$\boldsymbol{\beta}_j = \mathbf{W}_j\boldsymbol{\gamma} + \mathbf{u}_j \tag{1.2}$$

Each $\mathbf{W}_j$ has dimension $p \times q$ and is a matrix of background variables on the $j$th group and $\mathbf{u}_j \sim N(0, \boldsymbol{\tau})$. Clearly, since $\boldsymbol{\tau}$ is not necessarily diagonal, the elements of the random vector $\boldsymbol{\beta}_j$ are not independent.

Combining equations yields the single equation model:

$$\mathbf{Y}_j = \mathbf{X}_j\mathbf{W}_j\boldsymbol{\gamma} + \mathbf{X}_j\mathbf{u}_j + \mathbf{r}_j, \tag{1.3}$$

*Key Words and Phrases:* NELS:88, prediction, multilevel model, cross-validation, bootstrap

* School of Business Administration, University of Miami

** Department of Statistics, University of California, Los Angeles

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

[1)] Viewing equation (1.1) 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).

which may be viewed as a special case of the mixed linear model, with fixed effects $\boldsymbol{\gamma}$ and random effects $\mathbf{u}_j$.[2] Marginally, $\mathbf{y}_j$ has expected value $\mathbf{X}_j\mathbf{W}_j\boldsymbol{\gamma}$ and dispersion $\mathbf{V}_j = \mathbf{X}_j\boldsymbol{\tau}\mathbf{X}_j{}' + \sigma^2\mathbf{I}$. Observations in the same group have correlated disturbances, and this correlation will be larger if their predictor profiles are more alike in the metric $\boldsymbol{\tau}$ (De Leeuw & Kreft, 1995). Thus, the full log-likelihood for the $j$th unit is

$$L_j(\sigma^2, \boldsymbol{\tau}, \boldsymbol{\gamma}) = -\frac{n_j}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}_j| - \frac{1}{2}\mathbf{d}_j'\mathbf{V}_j^{-1}\mathbf{d}_j, \tag{1.4}$$

where $\mathbf{d}_j = \mathbf{Y}_j - \mathbf{X}_j\mathbf{W}_j\boldsymbol{\gamma}$. Since the $J$ units are independent, we write the log-likelihood for the entire model as a sum of unit log-likelihoods, *i.e.*,

$$L(\sigma^2, \boldsymbol{\tau}, \boldsymbol{\gamma}) = \sum_{j=1}^{J} L_j(\sigma^2, \boldsymbol{\tau}, \boldsymbol{\gamma}). \tag{1.5}$$

In an earlier study, we consider the problem of predicting a future observable $y_{*j}$ in the $j$th group of the framework above (Afshartous & de Leeuw 2004). We examined various prediction rules for predicting $y_{*j}$ and demonstrated several analytical results on the relative performance of these prediction rules.[3] Moreover, the multilevel prediction approach employing a shrinkage estimator for the regression coefficients within each group proved to be the most accurate over a wide range of simulations.[4] Recall that this shrinkage estimator may be written as a linear combination of the OLS and Prior estimator of the group regression coefficients (see e.g. Bryk & Raudenbush, 1992, p.43). The multilevel prediction rule for the future observable is simply a linear function of the shrinkage estimator of the level-1 regression coefficients for the particular group.

In this article, we apply the multilevel prediction approach to a real data set to assess predictive performance in practice. In addition, we develop model selection criteria (multilevel cross-validation and multilevel bootstrap) that distinguish the predictive ability of candidate multilevel models. We introduce two plots that 1) aid in the visualization the amount to which the multilevel model predictions are "shrunk" or translated from the OLS predictions, and 2) help identify if certain groups exist for which the predictions are particularly good or bad.

The multilevel prediction approach is applied to a portion of the base-year sample from the National Educational Longitudinal Study of 1988 (NELS:88). The base-year sample consists of 24,599 eighth grade students, distributed amongst 1052 schools nationwide. Given the plethora of student and school level variables available from the NELS:88 data (over 6,000), an endless number of multilevel models may be proposed and estimated. We

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

[3] Specifically, we investigated the behavior of the various prediction rules as 1) the number of units per group increases, and 2) as the intraclass correlation varies. The intraclass correlation measures the degree to which observations within the same group are related.

[4] Accuracy was assessed via the popular measure of predictive mean square error. Moreover, the simulation study extensively covers both the sample size and parameter space. Specifically, the sample size space concerns the various combinations of level level-1 (individual) and level-2 (group) sample sizes, while the parameter space concerns different intraclass correlation values.

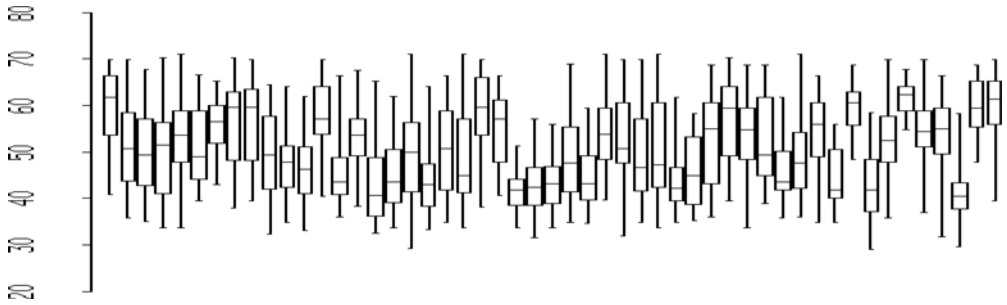


Figure 1: Distribution of mathematics score in 50 schools from NELS:88 data

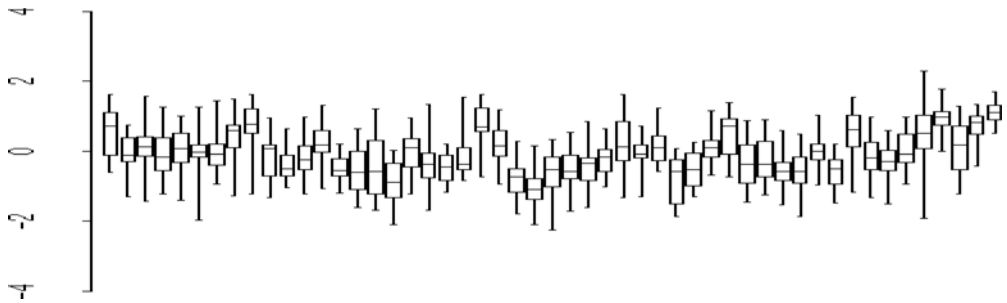


Figure 2: Distribution of SES in 50 schools from NELS:88 data

consider student mathematics score modeled as a function of the race and socio-economic status (SES) of the student, while schools are differentiated according to school type (public versus non-public) and the extent to which school lunches are subsidized.[5)] A stratified random sample of 50 schools was taken from the 1052 schools, yielding a total of 1152 students.[6)] There are 19 to 64 students within each school with a median of 24 students per school. Students for which mathematics score data was missing were removed from the sub-sample. The average mathematics score for the 1152 students in the sample is 50.7590 with a standard deviation of 9.9683; The variability of mathematics score from school to school is displayed in Figure 1. The SES variable is a standardized variable and thus has an average of -0.0641 and a standard deviation of 0.7416; The distribution of SES from school to school is displayed in Figure 2. In section 2 we propose four candidate multilevel models for the NELS:88 data set; we estimate each of the four models and compare and contrast the adequacy of each model. In section 3 and 4 we propose and apply cross-validation and bootstrapping algorithms, respectively, to assess the predictive per-

[5)] The extent to which school lunches are subsidized may be considered as an indicator of the poverty level of the students within a given school. The whites/Asians variable was formed as a composite variable from the respective indicator variables for whites and Asians.

[6)] The stratification was with respect to the public/non-public school dichotomy, with the percentage of each being set equal to that in the populations. This method resulted in 39 public schools and 11 non-public schools. The non-public schools consist of private, Catholic, and other religious schools.

formance of each of the four models; in section 5 present a brief summary and directions for future research.

## 2. Four Models

Using the two level-1 and two level-2 variables mentioned previously, we consider four candidate models for our sub-sample of 50 schools. Model 1 employs student SES as the level-1 variable and Public as the level-2 variable, where public is a 0-1 indicator variable with 1 representing public schools. Model 2 is the same as Model 1 but for the fact that the level-2 variable is G8Lunch, a categorical variable indicating the degree to which school lunches are subsidized in the school; this variable may be interpreted as a proxy variable for the overall poverty level of the school. Model 3 is the same as Model 1—with SES as the level-1 variable and Public as the level-2 variable—except that a level-1 whites/Asians variable is introduced, coded 1 for whites/Asians and 0 for all others. Model 4 is the same as Model 3 except the level-2 variable is G8Lunch instead of Public. The estimation results for each of these models are presented below. Note that in each case the full model is estimated, i.e., all the coefficients are random and there is a covariance(s) between the coefficients. After the presentation of the estimation results, we shall compare and contrast the estimation results for the different models. Table 1 below lists the level-1 and level-2 variables for each of the models:

Table 1: Four Candidate Models

| Model | level-1 | level-2 |
|---|---|---|
| Model 1 | SES | Public |
| Model 2 | SES | G8Lunch |
| Model 3 | SES, Whites/Asians | Public |
| Model 4 | SES, Whites/Asians | G8Lunch |

Formally, Model 1 may be written as follows:

$$\text{Math}_{ij} = \beta_{0j} + \text{SES}_{ij}\beta_{1j} + r_{ij}$$

where

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

Maximum likelihood estimates for Model 1 were produced by TERRACE-TWO and are given in Table 2.[7)]

[7)] 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 (1996) 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. Note: the log-likelihood for Model 1 may be directly obtained from equation (1.4) and equation (1.5). Each $\mathbf{X}_j$ has dimensions $n_j \times 2$, while each $\mathbf{W}_j$ has dimension $2 \times 4$ (block-diagonal) since we have two level-1 variables modeled as random.

Table 2: Model 1 for NELS:88 data

```
TERRACE-TWO:  Full Maximum Likelihood Estimates
                      Deviance     Method
Final Iteration  10:  8236.2026    Fisher

Parameters        Estimates     (S.E.)        T
Intercept
  By Intercept      51.7820   (  1.2357)    41.9064
  By Public         -0.9527   (  1.3895)    -0.6856
BYSES
  By Intercept       3.5715   (  0.9485)     3.7652
  By Public          0.9457   (  1.0655)     0.8875

Sigma^2:   69.2333
          Tau (covariance)

Intercept      12.6168   2.6826
BYSES           2.6826   1.1784

          Tau (correlation)
Intercept       1.0000   0.6957
BYSES           0.6957   1.0000
```

The output in Table 2 contains the estimates of the fixed effects for each of the level-1 random coefficients—the intercept and SES slope in this case. They may be interpreted as follows: The $\gamma_{00}$ estimate of 51.780 represents the expected mathematics score of a student attending a non-public school with a value of 0 for SES.[8)] Similarly, the $\gamma_{01}$ estimate of -0.9527 corresponds to the intercept shift between public and non-public schools. With respect to the SES-mathematics relationship, the estimate of 3.5715 represents the average value of this slope, i.e., a one unit increase in SES is associated with a 3.5715 unit increase in mathematics score on average, while the estimate of $\gamma_{11}$ corresponding to Public indicates a slightly steeper SES-mathematics relationship in public versus non-public schools. Also provided in Table 2 are the estimates of standard errors for these fixed effects and their corresponding t-ratios. Asymptotically we may treat these t-ratios in the usual manner, i.e., as normal z-values indicating the significance of the coefficient estimate (Bryk & Raudenbush, 1992). Table 2 also contains estimates of the level-1 and level-2 variance components, $\sigma^2$ and $\tau$ respectively. Note that since we have a level-2 variable, we may interpret the elements of $\tau$ as a conditional level-2 variance, i.e., the variation of the level-1 slopes that remains after accounting for certain level-2 variables. The legitimacy of modeling the level-2 slopes as random is illustrated via the overlayed added variable plots in Figure 3. We may interpret the separate lines as the OLS slope for the corresponding

[8)] Note that since SES is a standardized variable a value of 0 corresponds to the average score.

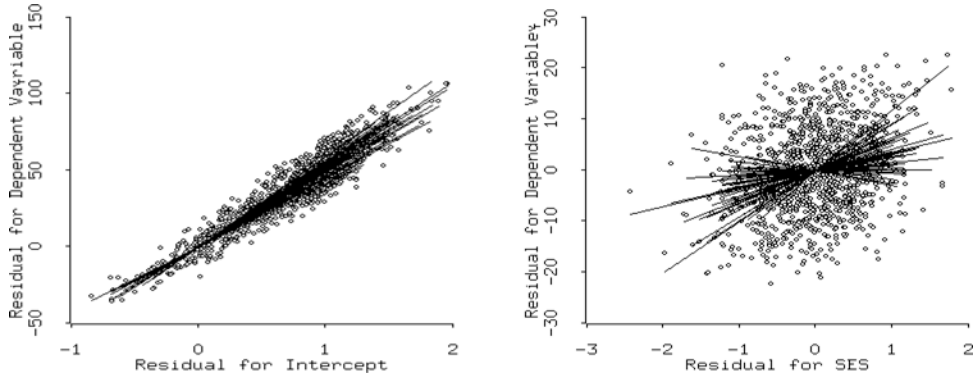


Figure 3: Model 1: Overlayed added variable plots for Intercept and SES for NELS:88 data

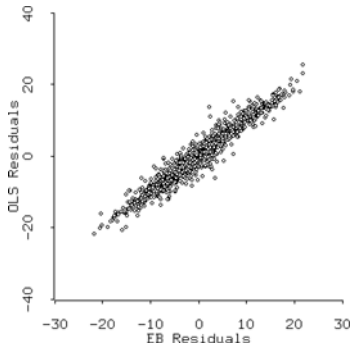


Figure 4: Double residual plot for Model 1

coefficient in the various schools. The variability of the SES-mathematics relationship is strong across the schools, while the intercept is not as variable. A useful diagnostic procedure for multilevel models is a "double residual" plot, where the residuals obtained from the multilevel estimation are plotted against those obtained from OLS estimation. Hilden-Minton (1995) introduced this double residual plot and suggested that such a plot is a good check of the level-2 model being entertained.[9)] Specifically, a well-fitting model should evince a strong linear relationship between both sets of residuals with a correlation approaching one. Figure 4 displays the double residual plot for Model 1; The correlation between the two sets of residuals is 0.969.[10)]

Formally, Model 2 may be written as follow:

$$\text{Math}_{ij} = \beta_{0j} + \text{SES}_{ij}\beta_{1j} + r_{ij}$$

where

[9)] The actual term double residual plot was provided by Jan de Leeuw.

[10)] Note that the residuals produced by the multilevel model are often referred to as empirical Bayes (EB) residuals.

Table 3: Model 2 for NELS:88 data

```
TERRACE-TWO:  Full Maximum Likelihood Estimates
         Deviance    Method
Final Iteration   9:  8227.9179    Fisher

Parameters        Estimates     (S.E.)        T
Intercept
  By Intercept      53.5417   (  0.9710)    55.1394
  By G8LUNCH        -0.8183   (  0.2559)    -3.1978
BYSES
  By Intercept       4.8650   (  0.7632)     6.3744
  By G8LUNCH        -0.2314   (  0.1996)    -1.1592
Sigma^2:   69.2205
          Tau (covariance)
Intercept      10.2139   2.0192
BYSES           2.0192   1.0362
          Tau (correlation)
Intercept       1.0000   0.6207
BYSES           0.6207   1.0000
```

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01}\text{G8Lunch}_j + u_{0j}\\ \beta_{1j} &= \gamma_{10} + \gamma_{11}\text{G8Lunch}_j + u_{1j}\end{aligned}$$

Maximum likelihood estimates for Model 2 were produced by TERRACE-TWO and are given in Table 3. The information in Table 3 may be interpreted in a similar manner to that of Table 2. The estimated models will be compared in the next section. Once again, the legitimacy of modeling the level-2 slopes as random is illustrated via the overlayed added variable plots in Figure 3.[11] Figure 5 displays the double residual plot for Model 2; the correlation between the two sets of residuals is 0.968, once again very high.

Formally, Model 3 may be written as follows:

$$\text{Math}_{ij} = \beta_{0j} + \text{SES}_{ij}\beta_{1j} + \text{Whites/Asians}_{ij}\beta_{2j} + r_{ij}$$

where

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01}\text{Public}_j + u_{0j}\\ \beta_{1j} &= \gamma_{10} + \gamma_{11}\text{Public}_j + u_{1j}\\ \beta_{2j} &= \gamma_{20} + \gamma_{21}\text{Public}_j + u_{2j}\end{aligned}$$

Maximum likelihood estimates for Model 3 were produced by TERRACE-TWO and are given in Table 4. The added variable plots and double residual plot cannot be produced for Model 3, since some of the schools possess singular level-1 design matrices and thus cannot produce OLS estimates.[12]

[11] We have the same added variable plots since the level-1 model is is unchanged.

[12] The singularity in the level-1 design matrices is a result of some of the schools having no variation

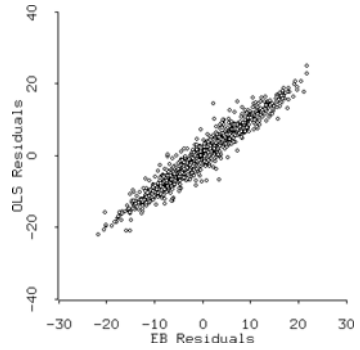


Figure 5: Double residual plot for Model 2

Table 4: Model 3 for NELS:88 data

```
TERRACE-TWO:  Full Maximum Likelihood Estimates
                      Deviance     Method
Final Iteration  11:  8194.6650    Fisher

Parameters        Estimates     (S.E.)        T
Intercept
  By Intercept      46.0834   (  1.6267)    28.3297
  By Public          1.2385   (  1.8191)     0.6808
Whites/Asians
  By Intercept       8.2003   (  1.8660)     4.3946
  By Public         -3.9427   (  2.0197)    -1.9521
BYSES
  By Intercept       3.2610   (  1.1523)     2.8299
  By Public          0.7585   (  1.2954)     0.5855
Sigma^2:   66.9927
          Tau (covariance)
Intercept       8.0358   1.1934   1.2787
Whites/Asians   1.1934   0.2324   0.6973
BYSES           1.2787   0.6973   5.5541
          Tau (correlation)
Intercept       1.0000   0.8733   0.1914
Whites/Asians   0.8733   1.0000   0.6138
BYSES           0.1914   0.6138   1.0000
```

Formally, Model 4 may be written as follows:

$$\text{Math}_{ij} = \beta_{0j} + \text{SES}_{ij}\beta_{1j} + \text{Whites/Asians}_{ij}\beta_{2j} + r_{ij}$$

in the whites/Asians variable.

Table 5: Model 4 for NELS:88 data

```
TERRACE-TWO:  Full Maximum Likelihood Estimates
                      Deviance     Method
Final Iteration  22:  8185.3118    Fisher

Parameters        Estimates     (S.E.)        T
Intercept
  By Intercept      48.5284   (  1.4064)    34.5043
  By G8LUNCH        -0.4304   (  0.3180)    -1.3534
Whites/Asians
  By Intercept       5.3982   (  1.4358)     3.7597
  By G8LUNCH        -0.2125   (  0.3354)    -0.6335
BYSES
  By Intercept       4.4992   (  0.9094)     4.9476
  By G8LUNCH        -0.2392   (  0.2366)    -1.0108
Sigma^2:   67.0669
          Tau (covariance)
Intercept       7.2827   1.3251   1.2838
Whites/Asians   1.3251   0.2613  -0.0315
BYSES           1.2838  -0.0315   4.4055
          Tau (correlation)
Intercept       1.0000   0.9605   0.2267
Whites/Asians   0.9605   1.0000  -0.0294
BYSES           0.2267  -0.0294   1.0000
```

where

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01}\text{G8Lunch}_j + u_{0j}\\ \beta_{1j} &= \gamma_{10} + \gamma_{11}\text{G8Lunch}_j + u_{1j}\\ \beta_{2j} &= \gamma_{20} + \gamma_{21}\text{G8Lunch}_j + u_{2j}\end{aligned}$$

Maximum likelihood estimates for Model 4 were produced by TERRACE-TWO and are given in Table 5. Once again the added variable plots and double residual plot cannot be produced since some of the schools possess singular level-1 design matrices and thus cannot produce OLS estimates.

There exist many ways in which these four candidate models may be compared. For instance, one may rely on the asymptotic properties of the t-ratios and favor models with more stable coefficient estimates. On the other hand, one may focus on the explained variance at level-1 and level-2, choosing the model that has the lowest variance components. As will be seen below, both of these methods quickly lead to confusion. Indeed, Snijders & Bosker (1994) have shown that the addition of explanatory variables in the multilevel model may actually lead to an increase in the estimates of level-2 variance components and have proposed two candidates for measuring the explained or "modeled variance" at

level-1 and level-2.[13] Nevertheless, we discuss these models in such a manner to illustrate the difficult problems encountered with respect to model selection with multilevel models with real data. Next, we take a predictive approach to model selection and examine how well these models perform predictively.

### *2.1 Model 1 versus Model 2*

Recall that Model 1 is identical to Model 2 but for the fact that Model 1 has Public as the level-2 variable and Model 2 has G8Lunch as the level-2 variable; both models have student SES as the level-1 variable. With respect to estimation, Tables 2–3 indicate that the estimates of the fixed effects of G8Lunch are more stable than those for Public, for both the fixed effects on the level-1 intercept and the fixed effects on the level-1 SES slope. Indeed, the t-ratios for Public are small in magnitude for both the intercept and and SES slope in Model 1. Furthermore, the intercept variance—$\tau_{00}$— is lower in Model 2 than in Model 1, indicating that the variability of the level-1 intercept from school to school is better explained by the level-2 variable G8Lunch than the level-2 variable Public; the SES slope variance—$\tau_{11}$—is similar in both models. The estimated level-1 variance is similar in both models and, as noted above, both models produce similar double residual plots. Thus, although there are similarities in the fit of these two models, it would appear that Model 2 is a better model for the 50 schools under consideration.

### *2.2 Model 3 versus Model 4*

Recall that Model 3 is identical to Model 4 but for the fact that Model 3 has Public as the level-2 variable and Model 4 has G8Lunch as the level-2 variable; both employ student SES and whites/Asians as the level-1 variables. Once again the estimates for the fixed effects of G8Lunch are more stable than those for Public, both for the fixed effects on the level-1 intercept and level-1 SES slope. However, with respect to the level-1 whites/Asians slope, the fixed effect for Public is more stable than that for G8Lunch, where both G8Lunch and Public have the effect of reducing the level-1 whites/Asians slope, although only weakly in the case of G8Lunch. In both models, the whites/Asians variable represents a positive level-1 intercept shift for average mathematics score. Both models produce similar estimates of level-1 and level-2 variance components. Thus, based on the level-1 and level-2 variance estimates, it is questionable which of these models is better.

---

[13] Briefly, their measures of modeled or explained variance at level-1 and level-2—${R_1}^2$ and ${R_2}^2$, respectively—are based on assessing the proportional reductions in mean squared prediction error. Although they show that population values of ${R_1}^2$ in correctly specified models becomes smaller when predictor variables are deleted, they also note that it cannot be proved in general that estimates of this quantity become smaller when predictor variables are added. Thus, the undesirable possibility of negative values for ${R_1}^2$ remains. Although the formulae for estimating ${R_1}^2$ and ${R_2}^2$ for random intercept models are straightforward, they note that transferring these results to the general multilevel model is rather tedious.

*2.3 Model 1 versus Model 3*

Although Model 1 and Model 3 both have Public as the level-2 variable and SES as the level-1 variable, Model 3 also uses the whites/Asians level-1 variable. The level-1 variance in Model 3 is noticeably lower than that of Model 1. With respect to the level-2 variance components, the estimated variance of the intercept increases and the estimated variance of the SES slope decreases with the addition of the whites/Asians level-1 variable between Model 1 and Model 3. The estimated variance of the whites/Asians slope is relatively small. Although the Public variable does not seem important in Model 1, in Model 3 it possesses a stable estimate with respect to the fixed effect on the level-1 whites/Asians slope, indicating that the effect of the whites/Asians variable is lower in public schools. With respect to the other level-1 variables, however, the level-2 Public variable is still weak. Although the level-1 variance is reduced in Model 3, we have the problematic result of an increase in the estimated variance for one of the level-2 coefficients.

*2.4 Model 2 versus Model 4*

Although Model 2 and Model 4 both utilize G8Lunch as the level-2 variable and SES as the level-1 variable, Model 4 also uses the whites/Asians level-1 variable. Once again, the larger model produces a lower estimate of the level-1 variance and mixed results for the level-2 variance components, where the estimated variance for the intercept increases and the estimated variance for the SES slope decreases. The G8Lunch fixed effect on the level-1 intercept is reduced in Model 4, both in magnitude and stability. In both models the G8Lunch effect on the SES level-1 slope is small in magnitude and stability. The whites/Asians variable in Model 4 seems to provide a better level-1 model.

In oder to improve upon the rudimentary and potentially problematic model comparisons above, we apply a predictive approach to these four models very similar to the simulations we employed to assess the performance of various prediction rules (Afshartous & de Leeuw, 2004). Specifically, instead of focusing on coefficient estimates we examine how well these models predict a future observable $y_{*j}$ in the $j$th group, e.g., a future student in a particular school. To be sure, we no longer have any future observables as in the simulations, for we have but one data set. Thus, we employ cross-validation and bootstrapping schemes in order to mimic the process of predicting future observables in an attempt to determine which of the various models produces the best predictions.

## 3. Cross Validation

Recall that we have 50 schools and a variable number of students per school. One could calculate the standard cross-validation leave-one-out estimate of prediction error similar to that from the general case (Stone, 1974). Recall that in the general case this involves the following calculation:

$$\text{CV} = \sum_{i=1}^{n} (y_i - \hat{y}_i^{(i)})^2 / n \tag{3.1}$$

where $\hat{y}_i^{(i)}$ represents the fitted value for observation $i$ that is computed with the $i$ th part of the data removed. For the multilevel model, this would entail summing $N = \sum_{j=1}^{50} n_j = 1152$ squared elements, where the prediction of each observation is based on the other $N - 1$ observations. To be sure, given the size and structure of the data set, this would produce a result very similar to that produced by the ordinary residual sum of squares where we do not perform a leave-one-out procedure. Moreover, this procedure would entail cycling through every observation within every group. Instead, we propose an approach that takes advantage of the data structure that need not cycle through every observation within every group. Specifically, the algorithm is as follows:

Multilevel Cross-Validation

1. For each of the $J$ schools, randomly select one student.
2. Estimate the multilevel model with the remaining $N - J$ students.
3. Form the corresponding predictions and average sum of squared errors for the J selected students;
4. Go back to step 1.
5. Stop after $m$ iterations and take the mean of the $m$ average sum of squared errors, i.e., the $m$ predictive MSEs.

This algorithm is similar in nature to that of our previous simulation studies (Afshartous & de Leeuw, 2004) where for each simulated multilevel data set an additional J observations were simulated as future observables. Here—since we no longer have the option of simulating data—we mimic this procedure by repeatedly selecting $J$ observations and, in essence, designating these as future observables. Several aspects of this algorithm are appealing. One, since the students to be predicted are selected at random, different combinations of students from the schools will be selected each time. For instance, although we may select a student from the same school more than once, the students that will be selected along with that particular student will certainly be different each time. Furthermore, by taking advantage of the data structure in this manner we may get an assessment of prediction error by cycling through the algorithm a relatively few number of times.[14] This is important since this reduces the computing time because new model estimates must be produced for each set of predictions. Indeed, how many times the algorithm is cycled can be varied by the user. Recall the well known combinatoric results that if there are $n_j$ students per school we have $n_1 n_2 \cdots n_J$ possible combinations of students. For our current data set with $J = 50$ this number is approximately $2.9 \times 10^{68}$. To be sure, the algorithm may be cycled fewer times than this.

We apply this group cross-validation method to the four candidate models in order

[14] Another possible approach to taking advantage of the data structure would be a leave-one-group-out approach instead of the leave-one-student-out-per-school approach. However, this approach presents serious problems with respect to the calculation of level-1 coefficient estimates in the school that is left out. Although a few approaches were investigated, none proved to be worth pursuing further. The leave-one-student-out approach probably does not leave out enough of the data while leave-one-school-out approach leaves out too much of the data.

to investigate which model(s) is best with respect to prediction. First, we calculate the average sum of squared errors for all observations that is produced by the four models when cross-validation is not employed. As in the simulation studies, we shall refer to this quantify as predictive MSE. It is well known that such an estimate of predictive error is too optimistic, the reason being that the sample that is used to form the parameter estimates is the same sample that is used to evaluate the predictive performance. Efron & Tibshirani (1993) refer to the latter as the "test sample" and the former as the "training sample." Table 6 displays the results for the four models, seeming to divide the performance of the models between the smaller (Models 1 and 2) and larger (Models 3 and 4) models, with the models within each group performing similarly from this perspective. With respect to predictive accuracy, the standard deviation of mathematics score across all of the schools is 9.96830, and the magnitude of the prediction error (the square root of the terms in the table) for all four models is less than this amount on average. To be sure, as with non-nested data, were the estimates from these models applied to predict future data, the prediction error would surely rise.

Table 6: Predictive MSE for the four models

| Model | Predictive MSE |
|---|---|
| Model 1 | 66.6080 |
| Model 2 | 66.6953 |
| Model 3 | 63.5978 |
| Model 4 | 63.7044 |

Table 7 displays the estimates of predictive MSE produced via the cross-validatory algorithm described above. The algorithm was iterated $m = 100$ times for each of the four models. As expected, these numbers are greater than the corresponding numbers in Table 6 since the observations being predicted were not used for model estimation. In addition, the results provide a differentiation between the models within the small and large model groups. Specifically, it now appears that Model 2 is preferable to Model 1. In other words, adding the level-2 variable G8Lunch to our model reduces the variability of the Math-SES relationship across schools. Although the t-ratios in the model comparisons of the previous section hinted at this, the variance components estimates were similar in both models. In addition, Model 3 now appears slightly better than Model 4, whereas before the results were ambiguous as well. In other words, when we have both SES and Whites/Asians as our level-1 variables, the addition of the Public variable to the level-2 model does not seem to help with respect to prediction. Figure 6 displays the distribution of predictive MSE over the 100 iterations of the algorithm, where we form the average squared error in predicting 50 randomly selected students each iteration. The side-by-side boxplots provide the additional information that Model 2—although not nearly the best with respect to its overall level of predictive MSE—clearly enjoys the advantage of being less variable over the iterations.

To be sure, these results are specific to our sub-sample of 50 schools from the entire NELS:88 data. It is reasonable to ask if the results of Table 7 apply to other sub-samples

Table 7: Cross validation PMSE for the four models

| Model | Predictive MSE |
|---|---|
| Model 1 | 74.2009 |
| Model 2 | 71.0282 |
| Model 3 | 70.4985 |
| Model 4 | 69.9195 |

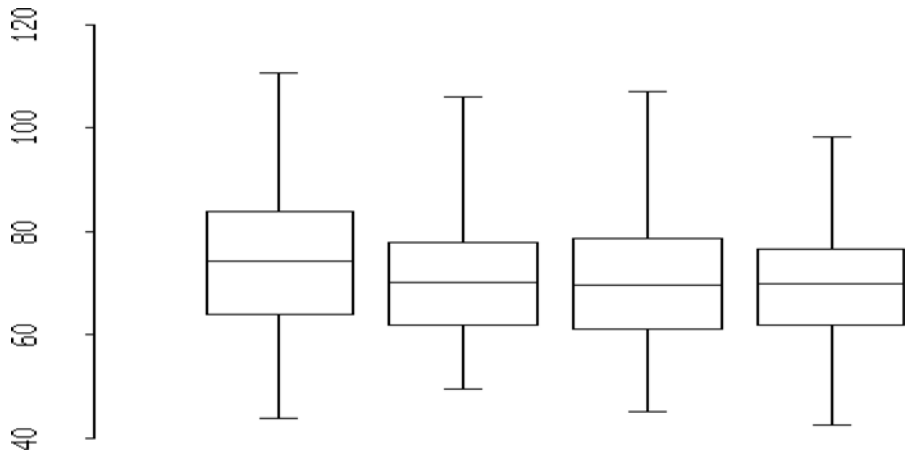


Figure 6: Distribution of Cross-validation PMSE for Models 1,2,3,4

of 50 schools. In order to investigate this question, we replicated the preceding cross-validation routine for 100 different sub-samples of 50 schools. Thus, for each of the 100 sub-samples of 50 schools, we iterate the cross-validation routine 100 times and calculate mean PMSE over these iterations. Table 8 below provides the overall mean cross-validation PMSE results taken across 100 different sub-samples of 50 schools, along with the relevant standard errors. The previous model ordering in Table 7 is maintained; however, due to the high variability of mean cross-validation PMSE across different sub-samples, we observe statistically significant differences only when comparing Model 1 to Model 3, and

Table 8: CV PMSE for the four models, averaged over 100 sub-samples

| Model | Predictive MSE |
|---|---|
| Model 1 | 72.6068 $\pm$ .3091 |
| Model 2 | 72.4598 $\pm$ .3069 |
| Model 3 | 71.7866 $\pm$ .3058 |
| Model 4 | 71.6623 $\pm$ .3199 |

when comparing Model 1 to Model 4.[15)] Thus, the hypothesis that the cross-validation algorithm may be slightly better with respect to distinguishing within the groups of models, based on the analysis of the initial sub-sample of schools, is not validated across other sub-samples of schools. Although the cross-validation routine may provide useful information for model selection for a specific group of schools, a larger sub-sample may be necessary to make inferences with respect to the population of schools in the NELS:88 data.

In the results presented thus far we have not been concerned with the predictive performance in the particular schools, as we have averaged over the 50 predictions of each iteration to produce a single estimate of predictive mean square error. However, the performance of the predictions according to school identification is often of interest. For interest, we might want to know if predictions are more reliable in some schools than others. Thus, we introduce two plots which may be formed from the output of the cross-validation algorithm which provide useful information about the particular groups. One plot provides information about the accuracy of prediction while the other provides information about the degree to which the multilevel model predictions differ from those produced via OLS.

We shall refer to the first plot as the *error plot*. The error plot illustrates the predictive accuracy within the groups over the iterations of the cross-validation algorithm. Instead of averaging the predictions for the 50 selected students each iteration, these values are retained and studied separately for each school. If one performs 100 iterations of the original algorithm one will have 100 predictions for each school and may examine the distribution of these predictions for the school. Figures 7–10 display the distribution of the prediction errors over the 100 iterations for each of the 50 schools for each of the four models; we have chosen to plot the errors instead of the squared errors to illustrate in which schools there is substantial over/under prediction occurring.

We shall refer to the second plot as the *translation plot*. The translation plot illustrates the degree to which the multilevel model predictions differ from those of OLS, i.e., predictions formed from separate OLS equations for each school. One may view the predictions formed via the multilevel model as predictions that are translated from the predictions which would have been formed via OLS. The difference between the multilevel prediction and the OLS prediction represents a measure of the amount of translation. Accordingly, one may examine the distribution of the translations over the 100 iterations of the cross-validation algorithm for each of the 50 schools. This illustrates which schools have their predictions translated the most. However, we only present this distribution for Models 1 and 2, since the OLS predictions were not obtainable for Models 3 and 4. These plots for Models 1 and 2 are presented in Figures 11–12. A particular school principle could

[15)] Specifically, we performed a paired 2-sample significance test for the difference in mean CV PMSE between models at the .05 level. We performed paired 2-sample tests since the same sub-samples of 50 schools were used in the cross-validation routine for the different models. The observed p-value for comparing Model 1 and Model 3 was .043, while that for comparing Model 1 and Model 4 was .025. The observed p-value for comparing Model 2 and Model 3 was .077. Note: the standard errors listed in Table 8 are for the individual models, not the difference sample relevant for the paired t-test.

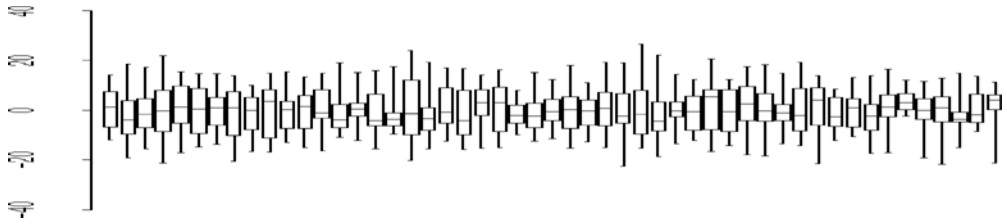


Figure 7: Model 1: Distribution of prediction errors in the 50 schools

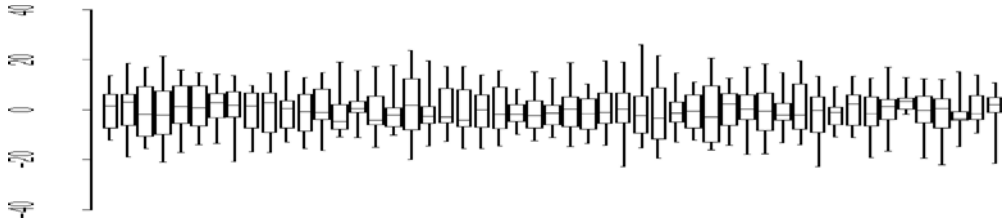


Figure 8: Model 2: Distribution of prediction errors in the 50 schools

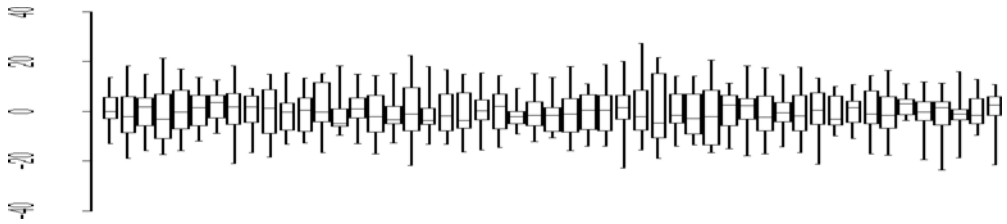


Figure 9: Model 3: Distribution of prediction errors in the 50 schools

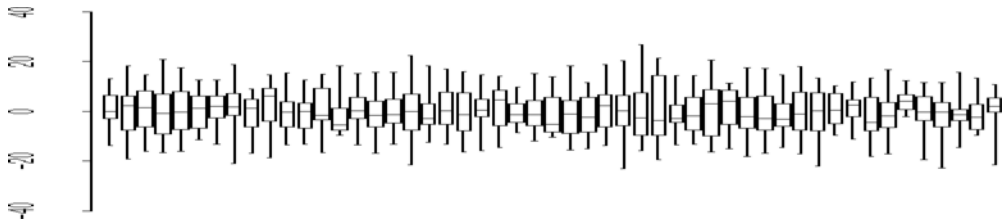


Figure 10: Model 4: Distribution of prediction errors in the 50 schools

investigate these plots to examine the amount of translation in the predictions for their school, both with respect to direction and magnitude.

## 4. Bootstrapping

A bootstrapping approach to assessing predictive performance may also be employed. Recall that in the previous section we first calculated the average sum of squared errors

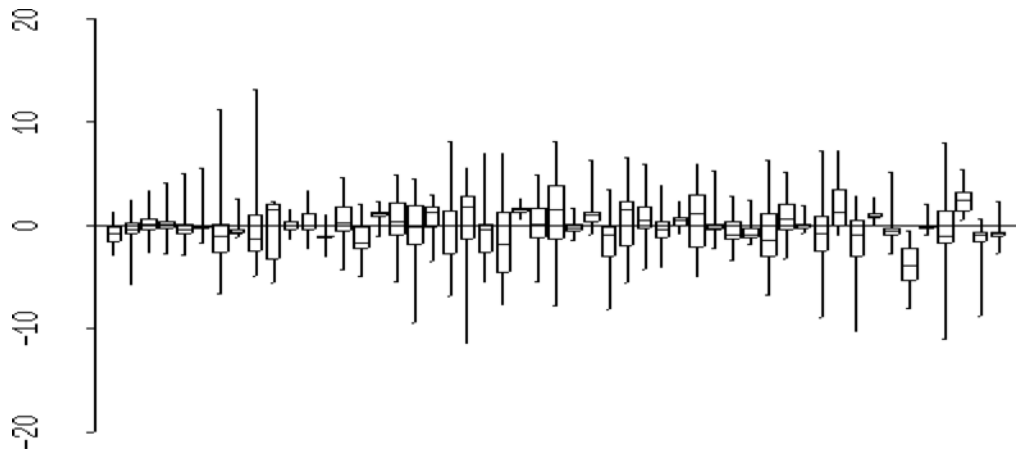


Figure 11: Model 1: Distribution of prediction translations in the 50 schools

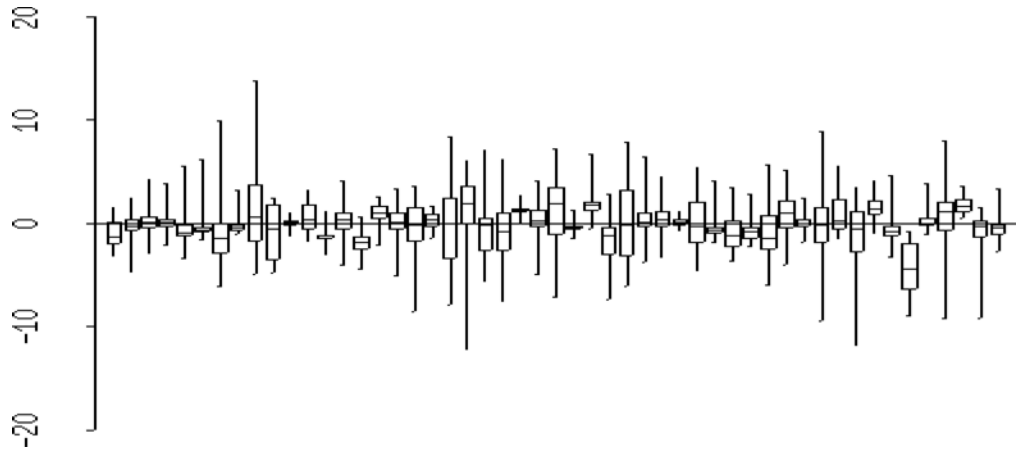


Figure 12: Model 2: Distribution of prediction translations in the 50 schools

produced by the four models on the original data, noting that this would be an underestimate of prediction error since the test sample and training sample are identical. The cross-validation algorithm was offered as a means of mimicking the predictive process that was studied in the simulations. Now, we offer the following bootstrap approach:

Multilevel Bootstrap

1. Set aside the original data.
2. Bootstrap the original data in the following manner: sample $n_j$ student with replacement from each of the $J$ schools.

3. Estimate the multilevel model using the bootstrap sample.
4. Form the predictions for all of the original data employing the coefficients obtained from the bootstrap sample.
5. Calculate the average sum of squared errors.
6. Go back to Step 2.
7. Stop after $m$ iterations and take the mean of the $m$ average sum of squared errors.

As with the cross-validation algorithm, this approach is appealing for several reasons. One, it better simulates the process of actual predictions since the bootstrap samples will be different from the original sample. Two, the user may alter the number of bootstrap samples. Indeed, given the structure of the data and the manner in which the bootstraps are taken, there exists a very large number of possible bootstrap samples even for data sets with a small number of groups.[16] Recall that for the cross-validation approach above there exists $n_1 n_2 \cdots n_J$ choices for the students to be predicted each iteration. Now, for the bootstrap approach the students to be predicted remain the same each iteration, while the sample producing the coefficient estimates changes each iteration.[17] Also note that for the cross-validation approach we are making only $J$ predictions each iteration while for the bootstrap approach we are making $N = \sum_{j=1}^{J} n_j$ predictions each iteration, i.e., the entire original data set is being predicted.

Table 9 displays the estimates of predictive MSE produced via the bootstrap algorithm described above. The algorithm was iterated 100 times for each of the four models. As expected, these numbers are greater than the corresponding numbers in Table 6 since we use bootstrap samples for estimation rather than the original data. With regard to model selection amongst the four models, the results of the bootstrap approach would lead to the same ordering of the four models as that from the cross-validation approach. For instance, Models 3 and 4—which contain the whites/Asians level-1 variable—clearly outperform Models 1 and 2 over the 100 iterations. The distinction between Models 1 and 2 and the distinction between Models 3 and 4 that was provided by the cross-validation approach, however, is now barely noticeable with the bootstrap approach, at least for this sub-sample of schools. Figure 13 displays the distribution of predictive MSE over the 100

Table 9: Bootstrap PMSE for the four models

| Model | Predictive MSE |
|---|---|
| Model 1 | 67.1643 |
| Model 2 | 67.1008 |
| Model 3 | 65.6405 |
| Model 4 | 65.4784 |

[16] Note that for data sets with few students per schools, this method runs the risk of producing bootstrap samples where some of the schools have singular level-1 design matrices, e.g., when the same student is selected $n_j$ times in a particular school for the bootstrap sample. Although this problem was encountered in previous simulations studies it was not a factor with the NELS:88 data.

[17] To be sure, the sample producing the coefficient estimates changes from iteration to iteration in the cross-validation approach as well, but in a much less corruptive manner than in the bootstrap approach since only $J$ observations are removed and the remaining $N - J$ observations are part of the original data.

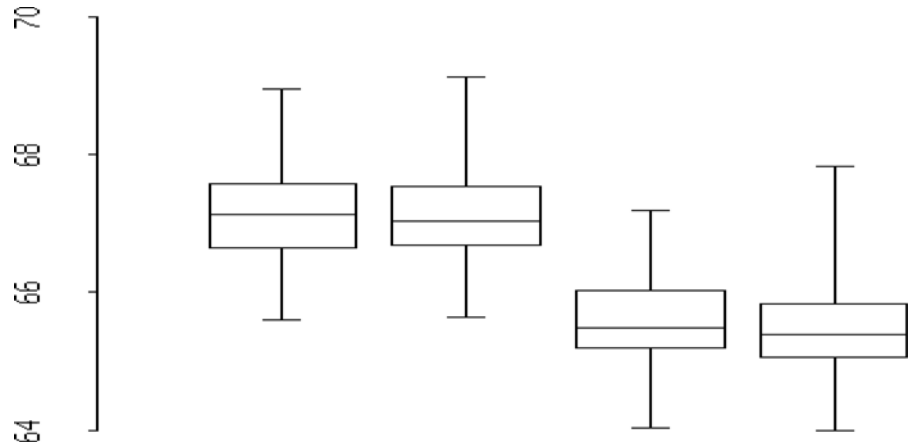


Figure 13: Distribution of Bootstrap PMSE for Models 1,2,3,4

iterations of the algorithm, where we form the average squared error in predicting the original data from a bootstrap sample each iteration. The lower variability in prediction for Model 2 in the cross-validation approach that was illustrated in Figure 6 does not appear in Figure 13 with respect to the bootstrap approach. However, Figure 13 illustrates how the distinction between Models 3 and 4 over Models 1 and 2 is now stronger.

As with the cross-validation results, it is reasonable to ask if the results of Table 9 apply to other sub-samples of 50 schools. In order to investigate this question, we replicated the preceding bootstrap routine for 100 different sub-samples of 50 schools. Thus, for each of the 100 sub-samples of 50 schools, we iterate the bootstrap routine 100 times and calculate mean Bootstrap PMSE over these iterations. Table 10 below provides the overall mean Bootstrap PMSE taken across 100 different sub-samples of 50 schools, along with the relevant standard errors. The previous model ordering in Table 9 is maintained, and we observe statistically significant differences for all pairwise model comparisons, except for comparing Model 3 versus Model 4.[18)]

Table 10: Bootstrap PMSE for the four models, averaged over 100 sub-samples

| Model | Predictive MSE |
|---|---|
| Model 1 | 67.9035 ± .2850 |
| Model 2 | 67.8592 ± .2830 |
| Model 3 | 65.8672 ± .2774 |
| Model 4 | 65.8438 ± .2797 |

[18)] We performed a paired 2-sample significance test for the difference in mean Bootstrap PMSE between models. The observed p-values were all statistically significant at the .01 level, except for the Model 3 versus Model 4 comparison, where the observed p-value was .053. We performed paired 2-sample tests since the same sub-samples of 50 schools were used for the different models. Note: the standard errors listed in Table 8 are for the individual models, not the difference sample relevant for the paired t-test.

Next, we consider a minor alteration to this bootstrapping method in order to illustrate an approach to a more specific prediction problem. Suppose we are interested in a fixed group of students from our original data set of fifty schools; say we are interested in one student per school and thus 50 students. The predictive performance of the four respective models may be examined with respect to predicting these particular 50 students. The bootstrap algorithm used to investigate this problem is identical but for the fact that the predictions are made only for these fifty students each iteration and these fifty students are not available for the bootstrap samples. Formally, the algorithm is as follows:

1. Randomly select one student per school and remove these students from the original data. These are the 50 students of interest; one may view them as 50 incoming students for which predictions are desired.
2. Estimate the model on the remaining $N - J$ observations and form the predictions and corresponding squared errors for the 50 students and take the average of these 50 squared errors. This represents the initial estimate of predictive MSE for these students.
3. Bootstrap the $N - J$ observations in step 2 and repeat step 2. This is a bootstrap estimate of predictive MSE for these 50 students.
4. Stop after $m$ iterations of Step 2 and take the average of the $m$ estimates of predictive MSE. Compare this to the result of Step 2.

Table 11 displays the initial estimates of predictive MSE for a randomly selected group of 50 students from the 50 schools.[19] Once again the models separate into two groups, but notice that for these particular students Model 4 outperforms Model 3 slightly, at least based on this one data set. Table 12 displays the estimates of predictive MSE for these 50 students produced via the bootstrap algorithm described above. The algorithm was iterated 100 times for each of the four models. With respect to model selection, the ordering of the models is now restored, where we prefer Model 4 over Model 3. This demonstrates the caution that should be applied to the predictions from the original sample, even if the observations to be predicted are omitted from this original sample.

Table 11: Initial estimates of PMSE for 50 students

| Model | Predictive MSE |
|---|---|
| Model 1 | 71.9280 |
| Model 2 | 71.1016 |
| Model 3 | 57.9067 |
| Model 4 | 58.0071 |

As with the cross-validation algorithm, the output from the bootstrap algorithm may be used to study different items of interest. For instance, instead of viewing the fixed

[19] The fifty students possessed the following characteristics: 35 of the students where white or Asian, the rest non-white/Asian; the average mathematics score was 52.4940, compared to a value of 50.8195 for the remainder of the data. The average SES of the 50 selected students was -0.0367, compared to an average SES of -0.06042 for the remainder of the data.

Table 12: Bootstrap estimates of PMSE for 50 students

| Model | Predictive MSE |
|---|---|
| Model 1 | 72.6800 |
| Model 2 | 72.3947 |
| Model 3 | 60.2810 |
| Model 4 | 60.0361 |

50 students as a group, one may be interested in predicting a particular student. Thus, instead of averaging the predictions for these 50 students each iteration, these values are retained and studied separately for each student. If one performs 100 iterations of the original algorithm, one will have 100 predictions for each student and may examine the distribution of these predictions for this student. Figures 14–17 display the distribution of the prediction errors over the 100 iterations for each of the 50 students for each of the four models; once again we have chosen to plot the errors instead of the squared error to illustrate for which students there is substantial over/under prediction occurring. To be sure, these results are somewhat specific to the students we have selected, but nevertheless the algorithm and resulting plot illustrates an approach to gauging the predictions for a

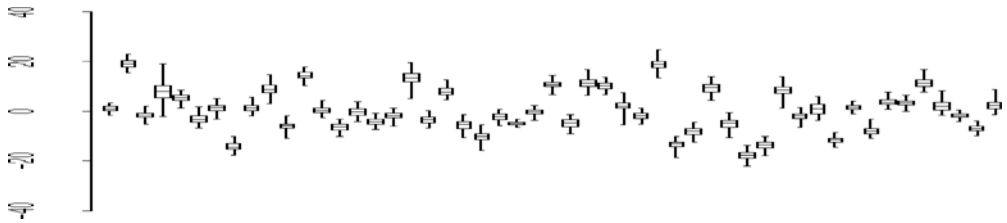


Figure 14: Model 1: Distribution of prediction errors for 50 students

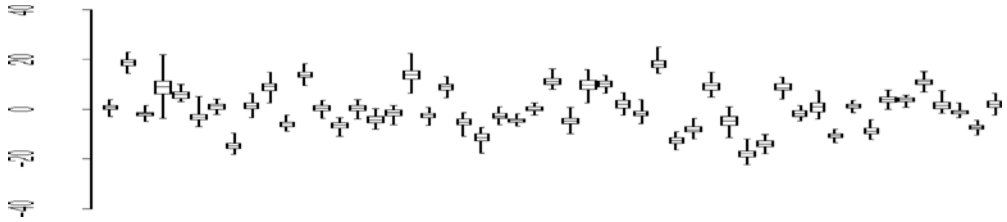


Figure 15: Model 2: Distribution of prediction errors for 50 students

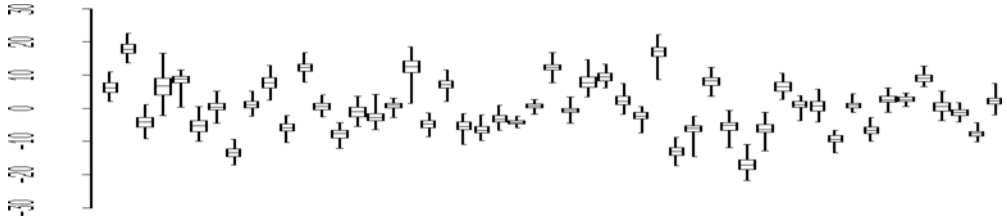


Figure 16: Model 3: Distribution of prediction errors for 50 students

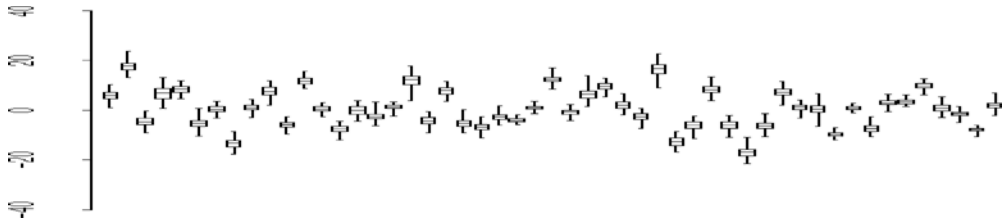

Figure 17: Model 4: Distribution of prediction errors for 50 students

particular student via bootstrapping.

## 5. Summary

We have applied the multilevel prediction approach to the NELS:88 educational data in order to assess its predictive performance on a real data set. We have seen that the addition of a level-1 variable reduces the predictive MSE for the models considered, as expected due to the reduction in conditional variability of the outcome to be predicted. For models with equal numbers of level-1 variables (e.g., Model 1 and Model 2), the distinction in predictive ability is often not as clear. We present model selection criteria, multilevel cross-validation and multilevel bootstrap, that are useful tools for assessing the predictive ability of candidate models. These cross-validation and bootstrap approaches often demonstrate distinctions in predictive performance that are often not as clear when examining only one sample and one set of estimates. In addition to applying these re-sampling routines to a sub-sample of the NELS:88 data, we investigate the stability of these results across repeated sub-samples from the NELS:88 data. These approaches adhere to the predictive perspective followed in our earlier study.

The two suggested plots, error plot and translation plot, capture several important aspects of the prediction process. Specifically, the error plot indicates whether certain schools exist in our NELS:88 sub-sample for which the predictions are particularly good or bad. The translation plot assesses how much the predictions of the multilevel method differ from those of the OLS method, i.e., how far the predictions are translated. These issues are especially salient from a substantive perspective. For instance, for graduate admissions data where one is attempting to predict the future performance of applicants, the use of translated or shrinkage predictors has the potential for controversy and thus requires careful study. To be sure, the NELS:88 data set considered in this article is but one data set; such methods need to be applied to many datasets before we can unequivocally endorse them.

## References

Afshartous, David, & de Leeuw, Jan (2004). "Prediction in Multilevel Models," forthcoming in *Journal of Educational and Behavioral Statistics*.

Afshartous, David, & Hilden-Minton, James (1996). "TERRACE-TWO: An XLISP-STAT Soft-

ware Package for Estimating Multilevel Models: User's Guide," *U.C.L.A Department of Statistics Technical Report*, www.stat.ucla.edu.

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.

Efron, E., & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall: New York.

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.

Leyland, A.H., & Goldstein, H. (2002). *Multilevel Modeling of Health Statistics*. New York: John Wiley.

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, 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.

Snijders, T.A.B., & Bosker, Roel, J. (1994). "Modeled Variance in Two-Level Models," *Journal of Education and Behavioral Statistics*, 22, pp.342–363.

Stone, M. (1974). "Cross-validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society, Series B*, 36, 111–147.

(Received December 2002, Revised June 2003)