Paper 365-2008

Applying SAS Mixed-Model Procedures to Educational Assessment Data

Brian F. Patterson, The College Board, New York, NY

ABSTRACT

Nested data structures that are most appropriately modeled using multi-level modeling techniques are common in educational research and just as commonly, the variable(s) of interest cannot be assumed normal a priori. The following study illustrates how a series of SAS® procedures may be used to investigate and evaluate each of these characteristics—normality of data and reasonableness of multi-level model. In this paper, this is managed by building a hierarchical generalized linear model (HGLM). The end goal of this research is to compare the within and across-group variance in exam performance for a national high school assessment program to inform possible educational intervention strategies.

SAS PROC KDE, 10-fold cross-validation and various tests of normality will be used to evaluate the distribution of the data and to inform the selection of the level-1 distribution and link functions in the HGLM setting. The results of these procedures and the substantial sample sizes allow the author to treat the data as approximately normal and fit a simpler HLM model using PROC MIXED. SAS experimental PROC GLMSELECT will be used in model specification for both univariate outcomes. Finally the research question will be addressed and SAS ODS Output and Graphics presented for a subset of the exams whose data that are suited to the proposed multi-level analysis.

INTRODUCTION

In educational research, analysts are often presented with nested data structures that may be most appropriately modeled using multi-level modeling techniques—also known as mixed models, hierarchical linear models, fixed- and random-effects models. A common example of nested data consists of students who are nested within classrooms, which are nested within schools which are nested within districts. Each of these levels could potentially have an impact on the any number of research questions. Also quite common when analyzing educational data, the dependent variable cannot be assumed normal a priori. If the outcome is a single score on an essay question with a score ranging from 1 to 6, then it cannot be presumed that this discrete outcome can be treated as normal. The study to follow illustrates how a series of SAS procedures may be used to account for both of these characteristics by building a hierarchical generalized linear model (HGLM) with the end goal of comparing the within-group and across-group variance in exam performance for a national high school assessment program for the two main sections of each exam. Readers interested in learning more about hierarchical linear modeling techniques are directed to the Suggested Reading section below.

BACKGROUND

The College Board's Advanced Placement Program® (AP®) was founded in 1955 as a national testing program with the goal of providing high school students opportunities to take college-level courses. As of the 2007 administration, the AP program consists of 37 courses (of which 34 have associated exams) in 22 different subject areas; the results from all 34 exams for the May 2007 administration are analyzed in this study. Of particular interest to many educational researchers is the comparison of the within- and across-school variance in academic performance. This is a particularly important question because the results of an analysis may impact policy decisions when developing interventions to improve academic performance. If much of the variance exists at the individual-level (that is, within group) then the appropriate intervention may be the remediation of lower-performing students. If however, a greater-than-expected proportion of variance is accounted for at the school-level (that is, across groups) then an intervention to inform teacher professional development may be in order for those low-performing schools.

Each test contains both multiple-choice, objectively scored, items and constructed-response or essay items that are scored by exam readers. Another element of the analyses to follow that may help refine the list of possible interventions is a careful comparison of variance in the objective-response (multiple-choice) section with the variance in the constructed-response (essay) section. For example, it may be that students demonstrate mastery of the content by doing well in the multiple choice section, but they fail to perform well on the free-response sections for some reason other than lacking expertise in the relevant subject-areas.

DATA

The data analyzed in the following sections are AP Exam data that include information both at the student-level such as student demographic and socio-economic measures and at the school-level such as school type (e.g., public or private). The dependent variables in this study are two-fold: (1) raw performance on objective-response (multiple-choice) items—referred to as the sum of unweighted objective parts or SUOP; and (2) raw performance on

constructed-response (essay) items—referred to as the sum of unweighted essay parts or SUEP. Data from the 2007 administration of all 34 AP Exams that have formal exams are analyzed both in separate univariate models for SUOP and SUEP and in a bivariate model that accounts for the covariance between these two measures. The exams differ in the number of multiple choice items and free response items that are administered.

EVALUATE DISTRIBUTION OF DEPENDENT VARIABLE

The first step in selecting the appropriate model—which will in turn dictate the appropriate SAS procedure to apply—is to review the graphs of the dependent data. In particular, Figures 1-2 and 3-4 show the empirical density functions of two multiple choice scores and two constructed response scores, using ODS graphics generated by SAS' PROC KDE. These are examples of the distributions that seem most in violation of the functional form of the normal distribution. ODS Graphics must be initialized and a device selected; Figures 1-4 were generated using the default GIF device because the vector-based enhanced metafile (EMF) device is not available as an ODS GRAPHICS output destination. The statement that initializes the ODS graphics capability is:

```
ODS GRAPHICS ON / IMAGEFMT= GIF IMAGENAME= "&Output_Dataset._&n._" RESET;
```

The relatively simple code used to generate graphs of the kernel density estimate is as follows:

```
PROC KDE DATA= &Output_Dataset._Lim;
   UNIVAR SUOP_Uncensored / PLOTS= DENSITY;
RUN;
```

And finally a few of the results of the above code are shown in Figures 1-4 below:

Figure 1. Empirical Distribution of SUOP: 2007 AP Calculus AB Exam.

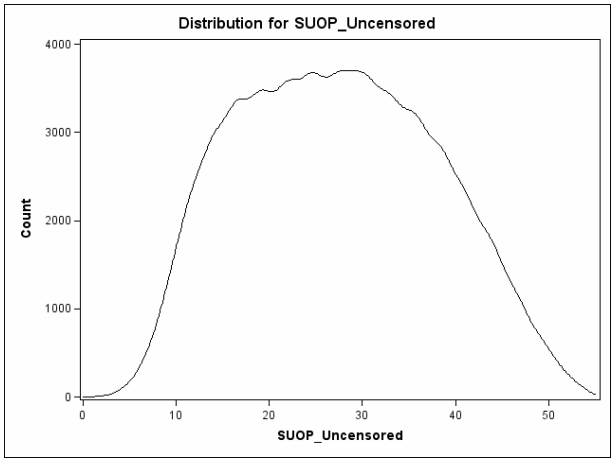


Figure 2. Empirical Distribution of SUOP: 2007 AP Physics B Exam.

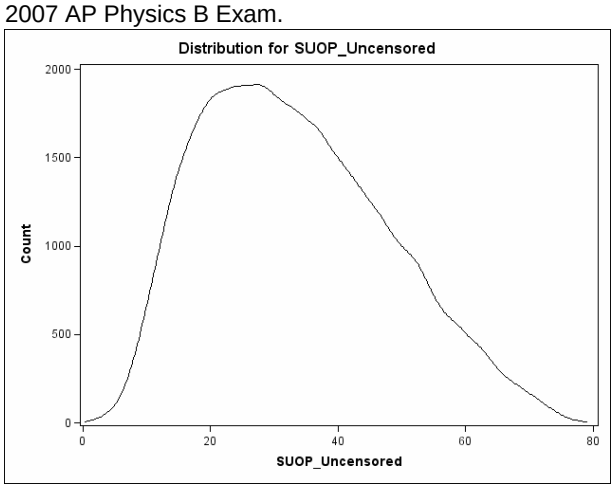


Figure 3. Empirical Distribution of SUEP: 2007 AP English Literature & Composition Exam.

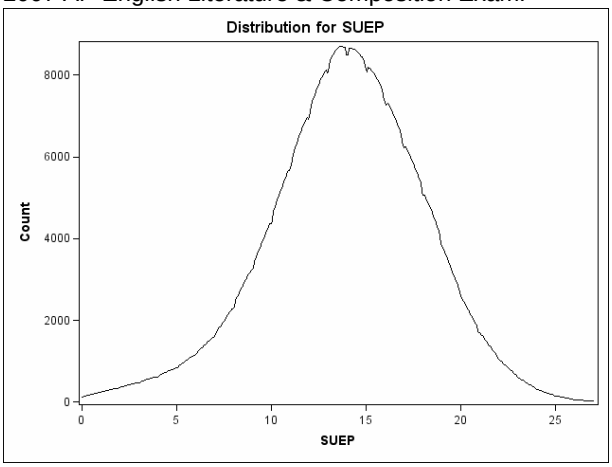
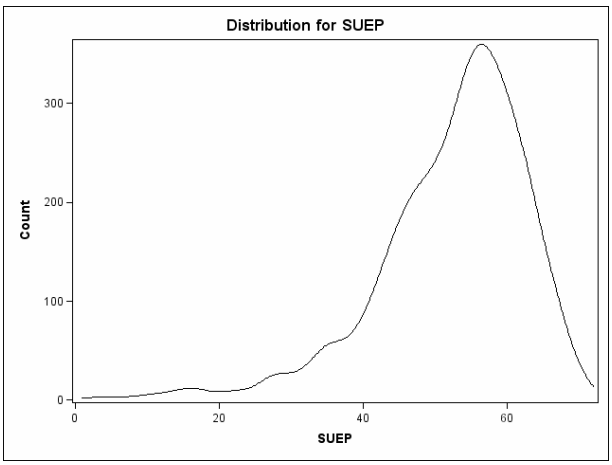


Figure 4. Empirical Distribution of SUEP: 2007 AP Chinese Exam.



The four empirical densities in Figures 1 through 4 show the two main issues that arise with the variables of interest: skew and kurtosis. Some exam parts (or sub-scores) appear normal—as does Figure 2—while others display just kurtosis—as in Figure 1—or in a few cases, both skew and kurtosis appear—as in Figures 3 & 4. It should be noted, however, that further analyses are necessary to determine if the apparent lack of fit to the normal distribution is such that a different analytic approach is called for. Some skew in the objective response composite has been reduced by choosing not to censor at zero (as is done in operational work). Even with such a correction, minor skew may persist for the objective-response score and there is no such restriction on the essay part sub-score, where skew is just a much an issue. Since no single problem affects all scores on all exams, we choose not to apply any variable transformations and instead rely on varying specifications of the level-1 distribution and link functions to address any non-normality in the dependent variables.

If the dependent variables were approximately normally distributed, then PROC MIXED may be appropriate—otherwise further research is required to determine whether PROC NLMIXED or PROC GLIMMIX would be optimal. After computing the hypothetical normal quantiles for each variable and comparing them to the observed, the next step was to perform 10-fold cross-validation by estimating empty PROC GLIMMIX HLM models for each of the distributions listed below. For details on how to easily and efficiently perform K-fold cross-validation, refer to Cassell (2007). In some cases the negative binomial distribution function outperforms the normal distribution in terms of root-mean-square-error (RMSE), but this tends to happen only for the smaller volume exams. 10- fold cross validation was performed on models for multiple-choice and constructed-response scores assuming each of the following level-1 distribution functions:

binomial;
negative binomial;
Poisson;
log-normal; and
normal.

Figure 5 below shows the results of the 10-fold cross-validation of the two exam sections of AP Environmental Science, where the optimal distribution—in terms of root mean squared error (RMSE) and in terms of mean absolute error (MAE)—for each section is asterisked for. In many cases, both RMSE and MAE are minimized under a normal distribution. Where the normal distribution was not optimal in terms of RMSE, the difference was usually so slight that the difference between the optimal and normal RMSE was less than 0.01—as is the case for the SUEP sub-score below where the negative binomial distribution is optimal in terms of RMSE, but the improvement beyond the normal distribution is so slight, that the author feels comfortable defaulting to the normal distribution in this case.

Figure 5. Example of 10-fold Cross-validation Results for AP Environmental Science.

Year	Dataset	Sub-Score	Distribution	Link	Observations	Root Mean Squared Error	Mean Absolute Error
2007	AP Env. Science	SUEP	Neg. Binomial	Log	48,670	6.3516 *	5.1144
2007	AP Env. Science	SUEP	Normal	Identity	48,670	6.3520	5.1176
2007	AP Env. Science	SUEP	Poisson	Log	48,670	6.3580	5.1130 *
2007	AP Env. Science	SUEP	Binomial	Logit	48,670	6.3625	5.1139
2007	AP Env. Science	SUEP	Log Normal	Identity	48,670	7.4276	15.2175
2007	AP Env. Science	SUOP	Normal	Identity	48,670	16.4367 *	13.2464
2007	AP Env. Science	SUOP	Neg. Binomial	Log	48,670	16.4534	13.2730
2007	AP Env. Science	SUOP	Poisson	Log	48,670	16.4580	13.2241 *
2007	AP Env. Science	SUOP	Binomial	Logit	48,670	16.4681	13.2266
2007	AP Env. Science	SUOP	Log Normal	Identity	48,670	19.7962	55.0591

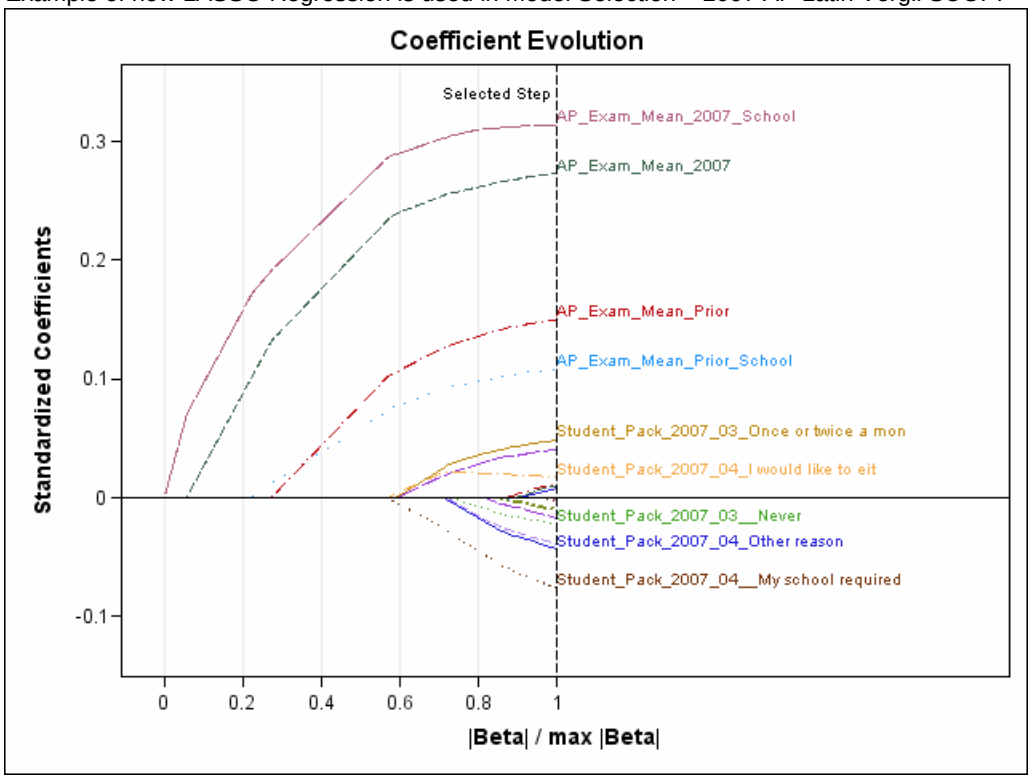
Performing formal tests for normality is the last check that was performed on the distributions of the dependent variables. With the NORMALTEST option in PROC UNIVARIATE, the results of a series of normality tests are generated. Reviewing the results of those tests, the author discovered that the issues initially identified in Figures 1 through 4 are not so severe that these data cannot be assumed normal.

Since the majority of these variables are approximately normally distributed, inference will be possible under the identity link function of traditional HLM. This is of particular interest in this case, since one of the stated goals of this research is to evaluate school- and student-level variance in order to potentially inform intervention strategies. If the exam data are approximately normal, then the intra-class correlation—that is the ratio of school-level (a.k.a. level-2) variance to the total variance—will be meaningful. In the case of non-linear level-1 link functions, this ratio is less meaningful because the variance at level-1 will be heteroskedastic (Raudenbush & Bryk, 2002, p. 298).

USE PROC GLMSELECT TO INFORM THE SPECIFICATION OF THE LEVEL-1 AND -2 MODELS

Since the focus of this research has to do mainly with prediction, this study uses Mallow's C(p) as the main model selection criterion and the method employed to specify the model is LASSO regression (Tibshirani, 1996). PROC GLMSELECT is run separately for the objective- and constructed-response sub-scores in order to evaluate whether important predictors are different for these two outcomes. When estimating the bivariate HLM model where the outcome is the pair of sub-scores, both sets of predictors for the marginal HLM models as selected above are included. For the sake of parsimony, the current study did not exhaust the possible interaction terms either within or across levels 1 and 2 of the HLM model, but further research may reveal the necessity for including additional terms.

Figure 6. Example of how LASSO Regression is used in Model Selection – 2007 AP Latin Vergil SUOP.



As is shown in this ODS GRAPHICS plot, at the optimal value of the shrinkage parameter for the LASSO model (see the dashed vertical line labeled "Selected Step"), the school-level mean AP exam scaled score for 2007 is the strongest positive predictor, confirming the author's expectation that school quality which is proxied by mean AP exam score for 2007 has a strong, positive relationship with student-level performance. The strongest negative predictor is the effect of an examinee indicating that their primary reason for taking the exam was that their school required him or her to take it. Note that the effect of the exam in question (i.e., AP Latin Vergil) removed from the student-level mean 2007 AP Exam score because the scaled score has a well-defined relationship to the objective-response section score—which is the outcome in this model. The other predictors have weaker relationships with the outcome, but are still included in this model that is optimal in terms of Mallow's C(p). One disadvantage to this graph is that SAS cannot label the effects that are clustered close together—see the unlabeled effects near where the horizontal axis intersects with the vertical dashed line. This effect can be identified by reviewing the ODS table called "SelectionSummary" or the ODS graph called "ASEPlot", which shows average squared error at as each effect is entered into the model.

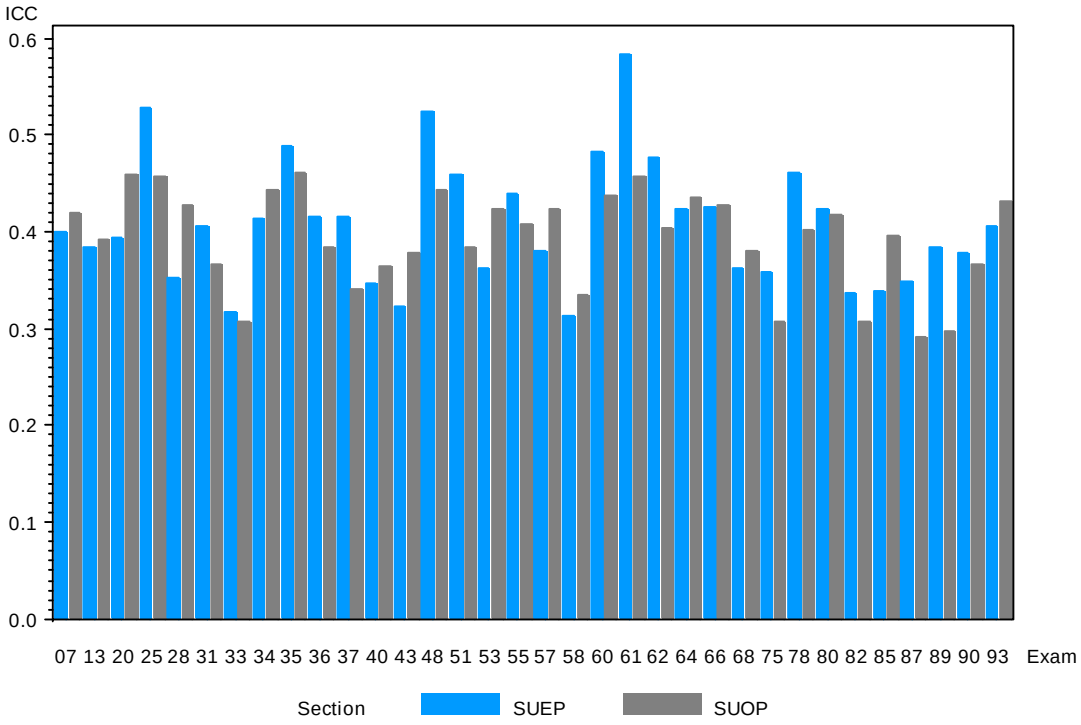
MODEL ESTIMATION & CHOICE OF SAS PROCEDURE

EVALUATE WHETHER HLM ANALYSIS IS NECESSARY

When the intra-class correlation (ICC) is relatively large, the usual interpretation is that a substantial portion of the variance in the outcome exists at level-2 or across-groups. In the cases where a linear level-1 link function is appropriate, the ICCs are computed and reported to establish the case for the appropriateness of a multi-level model. In the previous section, the author has demonstrated that the exam data are approximately normal the appropriate level-1 distribution and link functions are normal and identity, respectively.

The next element to examine from the output is whether the random effect terms associated with the level-2 equations are significant. If these terms are significant, then there is greater evidence for the need for the HLM approach. The intra-class correlations for each of the 34 exams for the objective- and constructed-response sections are computed by estimating an empty (i.e., intercept-only) HLM model with examinees at level-1 and school at level-2. Figure 7 shows a bar chart by exam (i.e., the two-digit exam code) and section of the intra-class correlations computed for each of the separate models estimated.

Figure 7. Plot of Intra-class Correlations for 2007 AP Exams by Section.



For both sections of the 34 exams in 2007, the high-school level variance term was highly significant ($p < 0.01$) and as Figure 7 shows, about 30 to 60 percent of the variance in these two scores is accounted for at the high school level. This finding supports the use of a multi-level modeling approach.

CHOOSE WHETHER TO USE PROC MIXED, PROC NLMIXED OR PROC GLIMMIX

If the distribution of the residuals and random effects are normal and the residuals are independent of the random effects, then the generalized linear mixed model may be replaced with a simpler mixed model, as in the case of this study. Other situations may call for a HGLM approach for which PROC NLMIXED and PROC GLIMMIX could be applied. As pointed out in Flom, McMahon & Pouget, (2007) other analysts should note that the "…[d]isadvantages of GLIMMIX are that the dependent variable has to be from an exponential distribution, whereas NLMIXED allows more flexibility (e.g. it can fit zero-inflated models), and that NLMIXED offers a true log likelihood, which GLIMMIX does not." In other words, if the skewness of the constructed-response sub-score—attributable to the censoring at zero—was substantial enough that the data could no longer be assumed normal, a zero-inflated approach as available with PROC NLMIXED may have been optimal.

SOME NOTES ON MANAGING MEMORY IN PROC MIXED

The bivariate HLM for some of the larger exams were so memory-intensive that the author's SAS script failed to execute. When that happened, the author found that SAS-L was an excellent resource for solving his problem and the following are some notes that he collected and implemented—some of which are operating-system-specific:

- If subjects are nested within the higher level unit—in this case students within schools—then specify this relationship in the REPEATED statement as such: REPEATED / SUBJECT= Reg_Number(High_School_Code) GROUP= Distribution TYPE= &Repeated_Cov_Type RCORR;.
- Note that the author noted substantial improvements in processing time for the bivariate model by changing the estimation method from REML to MIVQUE0.

- PROC MIXED requires contiguous memory when running and Windows XP users should note that there is a problem with service packs 1 and 2 that make estimation of large models with repeated terms difficult—if not impossible. For more information see SAS' Problem Note 15060 here: http://support.sas.com/kb/15/060.html.
- Another potential fix for Windows users—though it may result in some system instability—would be to add the following line to your SAS configuration file "-memsize 0". This option removes the pre-set limit on memory that SAS can use and should only be used as a last resort, after having tried to specify larger memory limits.

RESULTS

DISCUSS INTRA-CLASS CORRELATION ESTIMATES

Figure 7 shows that there is substantial (ICC > 0.30) and highly significant ($p < 0.01$) variance at the high-school level in the separate univariate models for each exam and section. Another finding of note is that the amount of variance that the high school level accounts for in the constructed response sub-score (SUEP) is generally greater than in the objective response sub-score (SUOP). The results of the analysis of intra-class correlations confirm the author's expectation that there is significant variance at the school-level in both objective-response and constructed-response exam performance. This result may support the use of teacher- or school-level intervention strategies as a tool for improving exam performance.

It should be noted that reliable data do not exist that would allow the identification of particular teacher or classroom effects within schools, which has two undesirable consequences: first, in the case where there are multiple teachers the effect of different teachers are all confounded with the variation due to the high school level; and on a related note, there may be some loss of power associated with test of significance for the regression parameter estimates as a result of the omission of the classroom- or teacher-levels [Moerbeek (2004)].

DISCUSS RESULTS OF UNIVARIATE MODEL ESTIMATES

A full discussion of the results from estimating each of the 68 univariate models is outside the scope of this paper, but one finding of particular interest for the author was the set of parameter estimates for a survey question regarding motivation. At the time of examination in 2007, AP examinees were asked, "What is your primary motivation for taking this AP Exam?" Across the 68 exams a consistent pattern arose: those students who claim to either "want to see how well [they] mastered the course material" or "would like to get college credit or exemption from introductory classes" significantly outperformed their peers who indicated that the primary reason for their having taken the exam was that "[their] school required [them] to take the exam." In educational research, the issue of motivation is always a challenge to operationalize and though this is not a perfect measure, it is more information that is available with many assessments.

DISCUSS RESULTS OF BIVARIATE MODEL ESTIMATES

Because of the substantial (generally $\hat{\rho} > 0.7$) and highly significant ($p < 0.01$) correlation between the two variables of interest, a more appropriate model would formally incorporate this fact and has in fact been estimated for each exam. The joint kernel density estimates for two exams are shown in Figures 8 & 9.

Figure 8. Bivariate Empirical Distribution: 2007 AP Music Theory Exam.

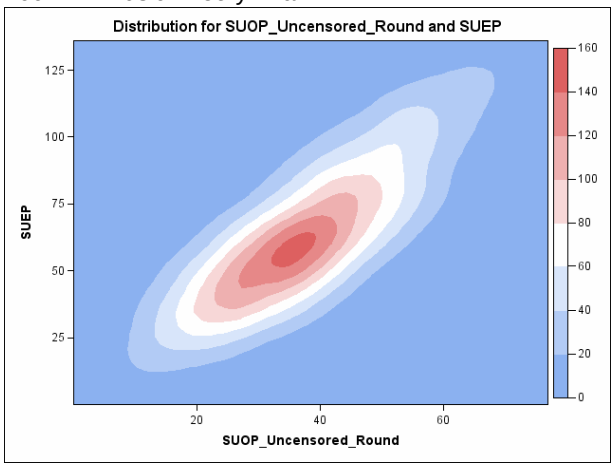
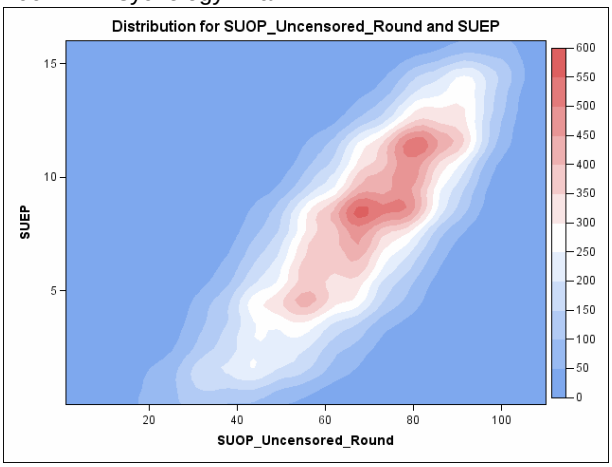


Figure 9. Bivariate Empirical Distribution: 2007 AP Psychology Exam.



Rather than a univariate outcome with one random effect for the intercept term associated with the high school, this model treats the two scores as repeated measures on the same individual. Such treatment complicates the covariance structure, because although both variables were shown to be approximately normal, they are not drawn from the same distribution. In an effort to account for this fact, the author first specified two RANDOM terms, one for the high-school level and one for the student level. Such an approach is supported by the significance of the school-level variance in the univariate model and the subsequent finding that within the bivariate model, there is significant covariance between scores on the two exam sections. The variance structure for the high school level would simply be the default variance components (VC) type, under which the covariance matrix is constrained to be a diagonal matrix (i.e., with no covariance terms between high schools). Clearly the covariance matrix at the student level should allow for the estimation of the covariance between objective response and constructed response section scores, so the variance structure in SAS' mixed-models terminology would be unstructured (UN). This covariance structure would be requested with the following PROC statement—note that the dataset contains two records per examinee, one for each section (SUOP and SUEP) and the "Distribution" variable identifies the section to which the record refers:

```
PROC MIXED DATA= &Dataset METHOD=MIVQUE0 IC LOGNOTE COVTEST;
   CLASS High_School_Code Student_Code Distribution;
   MODEL Joint_Response= Distribution / NOINT SOLUTION DDFM= RESIDUAL;
   RANDOM Distribution / SUBJECT= Student_Code(High_School_Code) TYPE= UN;
   RANDOM Distribution / SUBJECT= High_School_Code TYPE= VC;
RUN;
```

This author encountered difficulty in estimating the above model and instead ended up omitting the first RANDOM statement that corresponds to the high school level. The difficulties may be due to the fact that so much of the variation in the joint response is accounted for at the examinee level and because the two outcomes are so highly correlated. To compare the models, a variety of covariance structures, combinations of RANDOM and REPEATED statements and an effort to circumvent the problems by specifying a GROUP option on the RANDOM statements were used when estimating the above model. The model that was nearly always optimal in terms of the Akaike Information Criterion (AIC) was the above model without the RANDOM statement at the high school level.

DIRECTIONS FOR FUTURE STUDY

The author plans to continue to refine the covariance structure of the bivariate objective- and constructed-response multi-level model. In the interest of incorporating a measure of prior preparation into the model, the author plans to incorporate the exam results from the examinee's latest PSAT/NMSQT® score—which is generally taken in 10th or 11th grade. Finally, the author is considering a multi-level model that is cross-classified by high school and the examinees' zip codes. Such an approach may separate effects that appear to be related to the school and those related to the community in which the examinee lives.

CONCLUSIONS

The preceding study should serve as a practical case study in analyzing educational assessment data—or indeed data from any field—beginning with examining and specifying the distribution of the data through investigating any potentially multi-level aspects to the data, which inform the ultimate selection of the appropriate SAS procedure. Care should be taken in these early phases of the analysis, as they will naturally impact analysts' abilities to make either predictions or inferences based on the estimated model.

SUGGESTED READING

Texts such as Raudenbush and Bryk (2002) and Snijders and Bosker (1999) provide a good foundation for study in hierarchical linear modeling for those already familiar with traditional linear modeling techniques like ordinary least squares. Those already familiar with HLM who are interested in learning how to implement these models in SAS would benefit by reading Singer (1998) and those interested in estimating cross-classified HLMs could find useful guidance in Beretvas (2008).

REFERENCES

Beretvas, S.N. 2008. Cross-classified Random Effects Models. In Multilevel Modeling of Educational Data, A. A. O'Connell and D.B. McCoach (Eds.). Charlotte, NC: Information Age Publishing.

Cassell, D. L. 2007. "Don't Be Loopy: Re-Sampling and Simulation the SAS® Way." Proceedings of the 2007 SAS Global Forum. Cary, NC: SAS Institute Inc.

Cohen, R. A. 2006. "Introducing the GLMSELECT PROCEDURE for Model Selection." Proceedings of the 31st Annual Meeting of the SAS Users Group International. Cary, NC: SAS Institute Inc.

Ferrell, J. 2006. "A Comparison of the Mixed Procedure and the GLIMMIX Procedure." Proceedings of the 31st Annual Meeting of the SAS Users Group International. Cary, NC: SAS Institute Inc.

Flom, P. L., McMahon, J. M., and Pouget, E. R. "Using PROC NLMIXED and PROC GLMMIX to Analyze Dyadic Data with a Dichotomous Dependent Variable." Proceedings of the 2007 SAS Global Forum. Cary, NC: SAS Institute Inc.

Littell, R. C. 2007. "Repeated Measures Analysis with Clustered Subjects." Proceedings of the 2007 SAS Global Forum. Cary, NC: SAS Institute Inc.

Moerbeek, M. 2004. "The Consequence of Ignoring a Level of Nesting in Multilevel Analysis." Multivariate Behavioral Research, 39:1, 129-149.

Raudenbush, S. W. and Bryk, A. S. 2002. Hierarchical Linear Models: Applications and Data Analysis Methods. Second Edition. Newbury Park, CA: SAGE Publications.

SAS Institute. 2005. SAS Version 9.1.3 (executed on a Windows XP SP2 platform). Cary, NC.

SAS Institute. 2005. The GLMSELECT Procedure (Experimental). Cary, NC. http://www.sas.com/apps/demosdownloads/setupcat.jsp?cat=SAS%2FSTAT+Software.

SAS Institute. 2006. The GLIMMIX Procedure (Production). Cary, NC. http://www.sas.com/apps/demosdownloads/setupcat.jsp?cat=SAS%2FSTAT+Software.

SAS Institute. 2008. SAS OnlineDoc® 9.1.3, SAS Institute Inc., Cary, NC. http://support.sas.com/onlinedoc/913/.

Schabenberger, O. 2005. "Introducing the GLIMMIX Procedure for Generalized Linear Mixed Models." Proceedings of the 30th Annual Meeting of the SAS Users Group International. Cary, NC: SAS Institute Inc.

Singer, J. D. 1998. "Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models." Journal of Educational and Behavioral Statistics, Vol. 24, 323-355.

Snijders, T. A. B. and Bosker, R. J. 1999. Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling. London: SAGE Publications.

The College Board. "AP Central." http://apcentral.collegeboard.com/.

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso." Journal of the Royal Statistical Society Series B, 58, 267–288.

ACKNOWLEDGEMENTS

The author would like to acknowledge the valuable comments from his incredibly supportive colleagues at The College Board and those of Dr. Ann O'Connell.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Brian F. Patterson
The College Board, Research & Analysis
45 Columbus Avenue
New York, NY 10023
Tel: 212.713.7714
Fax: 212.649.8427
E-mail: bpatterson@collegeboard.org
Department Website: http://professionals.collegeboard.com/data-reports-research/