Paper 320-2012

Using SAS® PROC MIXED to Fit Health Policy-Based Hierarchical Models

Lori L. Miller, UC Davis Medical Center, Sacramento, CA
Fred Wilds, SAS Consultant, Seattle, WA

ABSTRACT

SAS® procedure PROC MIXED is a flexible procedure for fitting complex hierarchical linear models and calculating corresponding statistics. Documentation for PROC MIXED, however, remains complex and the defaults are often not appropriate. Using PROC MIXED does not preclude the need for substantial data processing to prepare for modeling, data analysis, and circumventing conversion issues due to the model including 36 public school groups and three or more student race/ethnicity and socioeconomic status categorical variables. This paper demonstrates: the process of aptly structuring a custom data set without use of array applications; creation of dummy coded variables for each categorical variable; and integration of the dummy coded variables, while omitting a chosen reference group, into PROC MIXED. Employing this approach expands the use of PROC MIXED.

INTRODUCTION

Health outcomes research is frequently multilevel in nature and presents researchers with conceptual, measurement, and methodological challenges (Lake, 2006a). Multilevel or hierarchical modeling coupled with the application of socio-ecological frameworks to the measurement plans and design of interventions are now more widely used in education and health behavior research (Elder et al., 2007; Singer, 1998). Multilevel analyses as such are statistical techniques (e.g., hierarchical linear modeling) that can simultaneously measure different levels of a hierarchy at individual (student) and group (school) levels (Lake, 2006b). Socio-ecological frameworks, however, are theoretical underpinnings characterized by multiple levels of influence on a behavior or health condition with an emphasis on capturing policy and environmental effects (Elder et al., 2007).

The hierarchical linear model, the main tool for multilevel analysis, is an extension of the standard linear regression model that is suitable for multilevel data (Snijders & Bosker, 1999). The SAS procedure PROC MIXED is a powerful yet flexible procedure integrated within the general purpose SAS statistical package (Base SAS 9.2) that allows for fitting very complex multilevel models and calculating all corresponding statistics. This program seeks to generalize the standard linear model allowing statistical analysis of mixed effects- fixed and random- for continuous outcome variables (Singer, 1998; Snijders & Bosker, 1999).

The scope herein is to highlight key aspects of data processing for using the SAS procedure Proc Mixed to analyze complex multilevel models in policy-based health outcomes research. The objective of this paper is to demonstrate how to aptly structure a custom data set without use of array applications, create dummy coded variables for each represented student race/ethnicity and socioeconomic status, and integrate the dummy coded variables for discrete or categorical variables, omitting a chosen dummy coded reference group, into the final mixed model procedure to enable smooth convergence of a two-level school health policy effects model.

THE TWO-LEVEL SCHOOL EFFECTS MODEL

The classic two-level school effects model, data having only two-levels within an organizational hierarchy, is the premise for the two-level school health policy effects model. This school effects model is one of the most commonly used hierarchical models in education research given that it is designed for data on individuals nested within naturally occurring hierarchies (e.g., students nested within schools). In this example of students nested within schools, level-1 predictors denote the individual students and level-2 predictors denote the schools as groups, illustrating the study focus on assessing the health behavior or condition of a level-1 outcome as a function of both level-1 and level-2 predictors (Singer, 1998).

Most policy-based health outcomes research traditionally requires the use of a large secondary data set. However, it is critical that a large hierarchically structured data set is selected for multilevel analyses. For this reason, the sample data used herein are drawn from a large secondary data set of the hierarchically structured (de-identified, restricted use) data from the 2003-2008 Trial of Activity for Adolescent Girls (TAAG) national research study. TAAG is a socio-

ecological theory guided randomized controlled multicenter field trial of 36 public middle schools at field sites in six states- Arizona, California, Louisiana, Maryland, Minnesota and South Carolina- as well as a coordinating center at the University of North Carolina funded by the National Heart, Lung, and Blood Institute, the National Institutes of Health. This research study assessed the effectiveness of a multicomponent school-based and community-linked intervention in preventing the decline in physical activity levels and cardiovascular fitness of racial/ethnically diverse adolescent girls, ages 12-14 years, from various socioeconomic statuses (SES). For the school health policy-based study discussed herein, the sub-sample population size was n=1336 public middle school girls in grades 6-8.

The statistical, fully unconditional, two-level school health policy-based model in Figure 1 below reflects the research question that addresses the effects of variation across school physical activity (PA) policy requirements [policy dose] for the dose-delivered in physical education (PE) class on change in students' body mass index (BMI) from 6^{th} to 8^{th} grade. In other words, BMI in 8^{th} grade (BMI8) is the primary outcome variable and a function of four level -1 and eight level -2 6^{th} grade level predictors.

- Level-1 comprises students BMI8 = BMI in 6^{th} grade (BMI6), age, race/ethnicity and SES.
- Level-2 characterizes the school context (school PA policy dose [of required PE minutes/week], PE dose-delivered [by PE teacher in class], PE class moderate-to- vigorous PA (MVPA) [student behavior response], enrollment size, percent minority enrollment, SES, and control or intervention group).

Note that both level-1 and level-2 predictors are fixed effects while the intercepts/slopes of students' BMI6 and the 36 public middle schools are random effects.

Figure 1. The Statistical Model for School Health Policy Effects:

Level-1

$$(StudentsBMI8)_{ij} = \beta_{0j} + \beta_{1j}(StudentBMI6)_{ij} + \beta_{2j}(StudentAge)_{ij} + \beta_{3j}(StudentRace/Ethnicity)_{ij} + \beta_{4j}(StudentSES)_{ij} + r_{ij}$$

Level-2

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(SchoolPAPolicyDose)_j + \gamma_{02}(SchoolPEDose\text{-}Delivered)_j + \gamma_{03}(SchoolPEClass\ MVPA)_j + \gamma_{04}(SchoolPEDose\text{-}Delivered*SchoolPEClassMVPA)_j + \gamma_{05}(SchoolPEClassSize)_j + \gamma_{06}(SchoolEnrollSize)_j + \gamma_{07}(SchoolPercentMinorityEnroll)_j + \gamma_{08}(SchoolSES)_j + \gamma_{09}(SchoolControl/Intervention)_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(SchoolPAPolicyDose)_j + u_{1j}$$

DATA PREPARATION STEPS:

Key Data Preparation Steps

As with all data sets, substantial data preparation is needed prior to data analysis. Using the SAS procedure Proc Mixed does not preclude this necessary process (Singer, 1998). Three key data preparation steps were needed prior to fitting the model and statistical data analysis:

1. Extract, transform, and load data from the secondary to the custom built data structure
2. Create dummy coded variables for student race/ethnicity and socioeconomic status (with > two categories)
3. Integrate dummy coded variables, omitting a chosen reference group, into the final mixed model procedure

Extracting variables from the secondary data set and transforming selected multiple data records to an appropriately structured custom built data set is the first key data preparation step. The next key data preparation step is to create dummy coded variables for all categorical variables particularly for those with more than two categories (e. g., student race/ethnicity and socioeconomic levels). Finally, after fitting both the empty and student model for which results are typically not interpreted or reported (Ma, MA, & Bradley 2008); integrate selected dummy coded variables, omitting the chosen reference group into the final mixed model procedure.

1. EXTRACT, LOAD AND TRANSFORM DATA FROM THE SECONDARY TO THE CUSTOM BUILT DATA STRUCTURE

After selecting all measurable variables of interest for the health policy-based outcomes research study from the selected hierarchically structured secondary data set, the next step is to aptly structure a custom built data set in preparation for analysis. Overall, the custom built data set should be structured as a wide to long multiple records data set for use in the SAS procedure Proc Mixed. The TAAG data required treating the data source as though you were extracting, transforming, and loading into the custom school health policy data set – TAAG INFO. For details on creating a SAS data set with the use of array applications in SAS procedure Proc Mixed refer to the appendix of Singer (1998).

Three-step process used to aptly structure the custom built data set:

- Extract the necessary data file sources by merging according to selected key variables
- Transform data into the necessary needed formats, which includes data cleansing, creation of new key variables, and creation of dummy coded variables for student race/ethnicity and socioeconomic status
- Load all totally transformed data into one data source to be analyzed by PROC MIXED

Approximately eighty percent of a researcher's time is spent solely on data preparation. The SAS code for extracting, transforming, and loading the measurable variables of interest from the secondary to the custom built data structure is more about data cleaning than actually creating an appropriately structured custom data set. Thus, an overview of the three-step process is provided along with emphasis on preparing the data extracted from several different data file sources for use in the SAS procedure Proc Mixed program for modeling and statistical analysis.

The secondary TAAG data was stored in several different data file sources in which we discovered many different issues that needed resolving. Data extraction and transformation began with a preliminary analysis of the data and data structure in terms of appropriateness for the research question, adequacy of quality data, and technical usability of the data and ended with reviewing the data codebook and dictionary, original questionnaire forms, and extracting appropriate variables to custom build a data set specific for this study.

Because data was being pulled from many different production data file sources, data extraction and transformation processing required gathering the data (measurable variables of interest) to load into one optimized single and unified custom built TAAG INFO data set that ensures providing a single source of data ready for analysis by SAS procedure PROC MIXED. These processes required identifying key variables on which to match and merge the measurable variables of interest to carefully renaming and formatting the values of the numeric variables to precision according the fields of nursing science and k-12 education. Note that it was necessary to carefully rename variables of interest so that we did not overwrite original data file sources in the secondary data set. In addition, these processes required creating new sequence numbers and key variables as needed along with evaluating the accuracy of all variables to ensure that there were no missing values for the numeric variables and to build a balanced custom data set.

Refer to Figure 2 for a schema of the process of extraction, transformation, and loading the measurable variables of interest from the six secondary data file sources - Policy dose data, Student data, School data, School lunch data, Policy Dose-Delivered data, and PE Classroom data - to aptly structure the custom TAAG INFO data set.

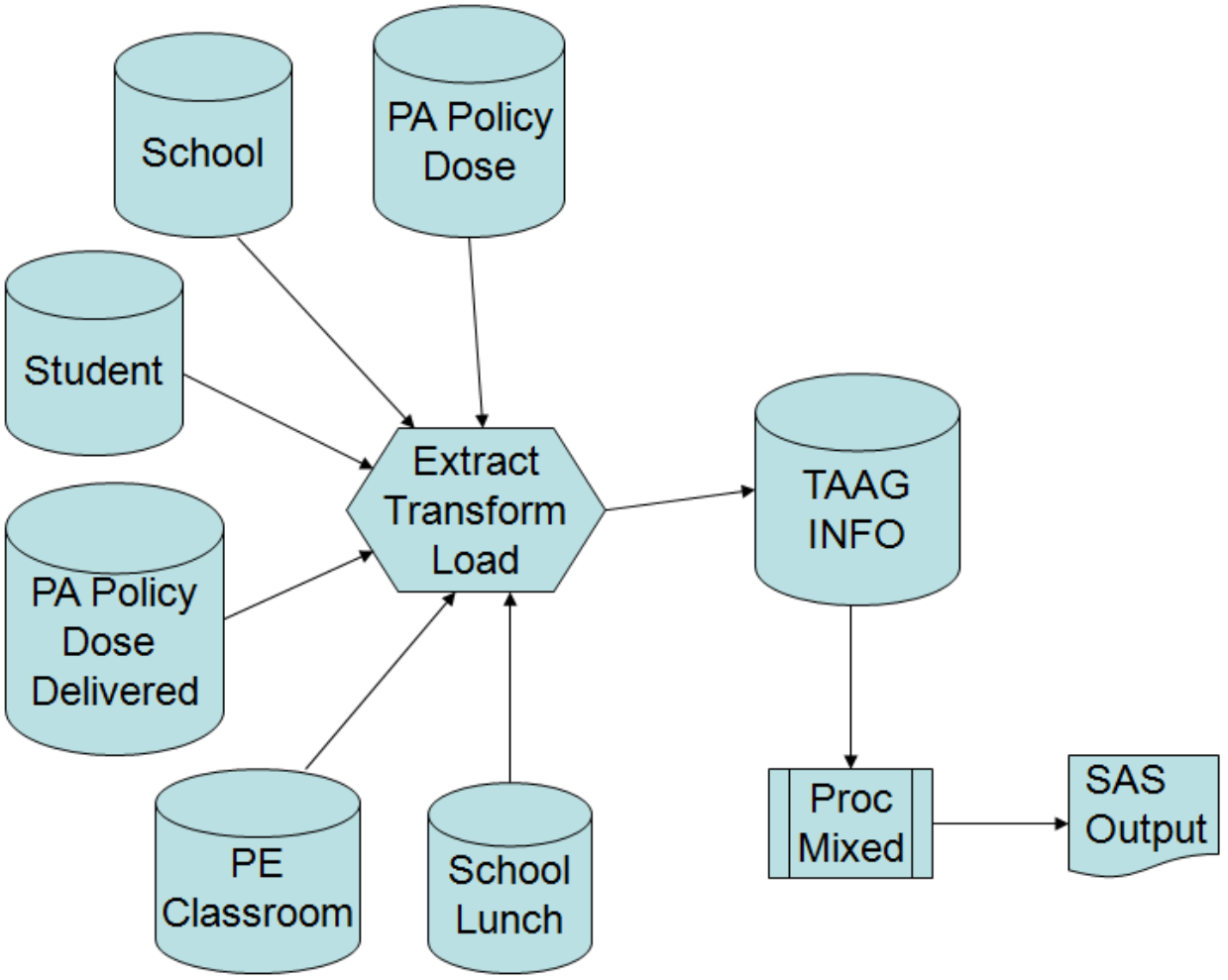

Figure 2. From Secondary to Custom Built Data Structure

2. CREATE DUMMY CODING CATEGORICAL VARIABLES FOR STUDENT RACE/ETHNICITY AND SOCIOECONOMIC STATUS

Model convergence issues due to the model including student race/ethnicity and socioeconomic status categorical variables with more than two categories are common in using PROC MIXED to fit complex hierarchical models. This particular convergence issue usually arises when more than 15 categorical variables, such as the 36 school groups are already included in the CLASS statement. Thus, as the model also included six student race/ethnicity and three socioeconomic status variables; placing these nine additional variables in the CLASS statement would not allow the model to converge. A simple work around to this problem is to dummy code each of the nine additional variables and later integrate each variable into the MODEL statement. This approach successfully enabled smooth converge while simultaneously drawing meaningful results.

Dummy coding is the most restrictive yet more commonly used coding scheme in regression models because it is the most easy to interpret. Student race/ethnicity and socioeconomic status are categorical variables with more than two categories that require the SAS user to carefully recode from categorical to continuous in order to enable the model to smoothly converge and draw meaning from the results. Dummy coding involves assigning values of “1” and “0” to reflect the presence or absence of a category (Gupta, 2008). And when there more than two or more uniquely defined categories to test, the researcher must choose a dummy coded reference group wherein all dummy coded variables are compared to the chosen dummy coded reference group, which is omitted from the final mixed model procedure.

To assess the change in BMI from 6th grade to 8th grade for a racial/ethnically diverse student population of six self-identifiable categories - Asian, Black or African American, Multiracial or Mixed Race, Spanish or Hispanic, Native American Indian or Alaskan Native, and White - create dummy coded variables for each represented race/ethnicity. In Program 1, note that the MI race/ethnicity category is the result of collapsing the Multiracial or Mixed race and Native American Indian race/ethnicity categories due to the very small number of students who self-identified as Native American Indian or Native Alaskan. Student SES level was categorized into three groups: yes (Y), no (N), and don't know (D). Since the schools were randomized to a control or intervention group in the original study, also included here are the dummy codes for categories: control (tmt_c) and intervention (tmt_i) treatment groups. After creating the dummy coded variables, be sure to evaluate the accuracy of the dummy coded variables by running a PROC FREQ to check whether the newly created variables contain any missing values. All dummy coded variables were created using the "IF/THEN" statement to assign binary codes of "1s" and "0s"as values. Program 1 illustrates this process. (Note that while more efficient methods are surely available, the code as illustrated in Program 1 was effective in meeting the needs of our project for the size of our database.)

Program 1. Creating Dummy Coded Variables

```
Data XYZ;
   set XYZ;

if race="A" then Asian=1;
if race in ("B","I","M","S","W") then Asian=0;

if race="B" then Black=1;
if race in ("A","I","M","S","W") then Black=0;

if race="I" then Indian=1;
if race in ("A","B","M","S","W") then Indian=0;

if race="M" then Multi=1;
if race in ("A","B","I","S","W") then Multi=0;

if race="S" then Spanish=1;
if race in ("A","B","I","M","W") then Spanish=0;

if race="W" then White=1;
if race in ("A","B","I","M","S") then White=0;

if race in ("I","M") then race_MI=1;
if race in ("A","B","S","W")then race_MI=0;

if studentses="1" then stses_y=1;
if studentses in ("2","3") then stses_y=0;

if studentses="2" then stses_n=1;
if studentses in ("1","3") then stses_n=0;

if studentses="3" then stses_d=1;
if studentses in ("1","2") then stses_d=0;

if tmt="C" then tmt_c=1;
if tmt="I" then tmt_c=0;

if tmt="I" then tmt_i=1;
if tmt="C" then tmt_i=0;

run;
```

3. INTEGRATE DUMMY CODED VARIABLES INTO FINAL MIXED MODEL PROCEDURE

Program 2 illustrates the SAS code mixed model template used for assessing the two-level school health policy effects model. Note that the level-1 and level-2 predictors are included in the MODEL statement while the RANDOM statement indicates the intercepts/slopes of BMI6 at the 36 schools are random.

Program 2. SAS Code Mixed Model Template

```
PROC MIXED DATA=XYZ METHOD=REML maxiter=5000 convh=1E-8 NOITPRINT NOCLPRINT;
CLASS SCH_ID;
MODEL BMI8 = BMI6 AGE MVPA ENROLL SCHOOLSES /solution;
RANDOM INTERCEPT BMI6 /SUBJECT=SCH_ID TYPE=UN;
RUN;
```

Integrate all student race/ethnicity and socioeconomic status dummy coded variables, while omitting the chosen dummy coded reference group, into the SAS mixed model template and run the model. Program 3 illustrates how each student race/ethnicity and socioeconomic status are now level-1 predictors that are integrated into the mixed model template while omitting the chosen dummy coded reference group of the researcher's choice - "white" race/ethnicity and student SES_N (no). Following each key data processing step as outlined enabled smooth convergence of the two-level school health policy effects model and meaningful results.

Program 3. SAS Code for Final Mixed Model Procedure

```
PROC MIXED DATA=XYZ METHOD=REML maxiter=5000 convh=1E-8 NOITPRINT NOCLPRINT;
  CLASS SCHOOL ID;
  MODEL BMI8 = BMI6 Age Asian Black Multi Spanish StudentSES_Y StudentSES_D
        MVPA School_Size School_SES Percent_Minority_Enroll PE_Class Size
        PE_Dose-Delivered*MVPA Policy_Dose Control_TMT_Group /solution ddfm=bw
        notest;
  RANDOM INTERCEPT BMI6/SUBJECT=SCHOOL ID type=UN;
RUN;
```

CONCLUSION

Using SAS procedure Proc Mixed allows the user flexibility in the data processing and statistical analysis of complex hierarchical linear models. This paper highlighted the importance of data preparation and the use of dummy coded variables through three key data preparation steps that enables smooth conversion of the SAS procedure Proc Mixed program to the two-level school health policy effects model: the appropriate set up of a custom built data set without the use of array applications; creation of dummy coded variables for each categorical variable; and integration of the dummy coded variables, while omitting a chosen reference group (for each represented student race/ethnicity and SES) into PROC MIXED.

Special purpose software programs (e.g., HLM) are specifically designed for applying multilevel modeling techniques. However, the HLM program, in particular, is not an integrated program within a general purpose statistical package as is the SAS Procedure PROC MIXED, for it requires the user to manage and process all data in SPSS, SAS, SYSTAT, or STATA prior to inputting separate files for each level of the design (Singer, 1998; Snijders & Bosker, 1999). In addition, there are currently no standard methods/approaches for policy-based health outcomes research. Employing the above outlined key data processing steps in preparing the data for modeling and analysis can expand the use of the SAS procedure PROC MIXED program mainly for research related to public health policy, public health nursing, school health policy, demographics, healthcare outcomes and, racial/ethnic health disparities.

REFERENCES

Elder JP, Lytle L, Sallis JF, Young DR, et al. (2007). A description of the social-ecological framework used in the trial of activity for adolescent girls (TAAG). Health Education Research, 22 (2), 155-165.

Gupta, R. (2008) Stat News#72: Coding categorical variables in regression models: Dummy and effect coding. Cornell University, Cornell Statistical Consulting Unit. Access on August 8, 2011 from http://www.cscu.cornell.edu/news/statnews/stnews72.pdf

Lake, ET (2006a). Multilevel models in health outcomes research, Part I: Theory, design, and measurement. Applied Nursing Research, 19, 51-53.

Lake, ET (2006b). Multilevel models in health outcomes research, Part II: Statistical and analytical issues. Applied Nursing Research, 19, 113-115.

Ma X, Ma L, & Bradley KD (2008). Using multilevel modeling to investigate school effects. In Multilevel Modeling of Educational Data. Information Age Publishing, pp. 59-110.

Singer, J (1998). Using SAS proc mixed to fit multilevel models, hierarchical models, and individual growth models. Journal of Education and Behavioral Statistics, 23 (4), 323-355.

Snijders, T & Bosker, R (1999). Multilevel analysis: An introduction to basic and advanced multilevel modeling. Sage Publications: Thousand Oaks, California.

ACKNOWLEDGEMENTS

The primary author is supported by a post-doctoral fellowship program from the Betty Irene Moore School of Nursing at University of California Davis Medical Center.

We extend a special thanks to our mentor Duke Owen, SAS Global Forum Mentoring Program and Rachael Biel and Tyler Smith, Statistics and Data Analysis Section. We also extend a special thanks to Julie Petlick, SAS Student Programs Manager - Education Division.

CONTACT INFORMATION

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

Lori L. Miller
Postdoctoral Fellow
Betty Irene Moore School of Nursing
University of California Davis Medical Center
4610 X Street, Suite 4202
Sacramento, CA 95817
Work Phone: (916) 889-6132
Email: loril.miller@ucdmc.ucdavis.edu

Fred Wilds
SAS Consultant
Excellence In Travel
14127 SE 182nd Street
Renton, Washington 98058
Work Phone: (425) 235-9175
Email: riofun3@comcast.net