Creation of an Educational Statistics Database at the University of Arkansas

Sean W. Mulvenon, University of Arkansas, Fayetteville, AR
Antoinette Thorn, University of Arkansas, Fayetteville, AR
Charles Aldrich, University of Arkansas, Fayetteville, AR

Abstract

The purpose of this project was to address the needs of the University of Arkansas in developing an educational database which could be made available for faculty, graduate students, and other community agencies requiring educational statistics and data. The first step in this process was the acquisition of data sets from the United States Department of Education (DOE) and the National Center for Educational Statistics (NCES). The second, was the writing of a program to create a managed database which would allow consolidation and merging of the data sets. This program was designed to work on any DOS operating system which can operate the SAS package and for users with intermediate expertise in using the SAS programming language. The SAS packages utilized were BASE, SAS/STAT, and IL.

Rationale

The University of Arkansas College of Education has traditionally emphasized graduate education leading to professional degrees, including a long-standing commitment to teacher preparation. In the last decade the University of Arkansas, as a land grant institution, has placed increased emphasis on research-based degrees and towards that end, programs have been expanded to include doctor of philosophy (Ph.D.) degrees. The shift in emphasis has created a renewed interest in developing and improving the quality and quantity of scholarly research produced in the College.

A goal of a quantitative research group based in the College of Education was to develop and maintain an educational database which could be made available for any faculty members, graduate students, and other community agencies which might require statistical analysis or data to conduct educational research. Ultimately, the availability of this database would facilitate acquisition and use of large national data sets which would improve the generalizability of results from studies traditionally relying on local data. Thus, the overall quality may be improved due to the amount of data and various constructs which could also be examined. The purpose of this paper is to describe the procedures and the use of SAS to complete this goal of creating and maintaining an educational database.

Method

The first step in developing this educational database was to locate data sets which could be obtained for unrestricted use without violating any copyright or private use restrictions. One source identified was the National Center for Educational Statistics (NCES), which is one of 13 divisions of the United States Department of Education (DOE). The National Center for Educational Statistics collects and maintains numerous large data sets on educational statistics which are available for public use for periods up to five years or longer if requested.

A possible question is why would it be necessary to generate this type of public database if data is currently available for public use? As previously mentioned, the intent was to simplify data access. If a database, consisting of the NCES data sets was maintained locally, then access by individuals requiring this data would be easier. Further, the data in these data sets were obtained using stratified sampling procedures. There are a number of statistical issues which are related to using these data sets as a result of the use of stratified sampling and include weighting of the variables and estimation of variance components. Experts at the NCES acknowledge that failure to address these statistical issues can lead to very spurious conclusions being drawn from research projects which utilize these data sets. A goal of creating this database is to provide support and assistance to graduate students and faculty members from the University of Arkansas who utilize the database.

Restricted Used Data

The NCES also maintains raw data for various data sets and this data is available under restricted use licenses to guarantee that subject anonymity is protected. All members of the quantitative research group have access to restricted data through the NCES and the application process for obtaining a restricted data license. The affidavit of nondisclosure precludes us from “giving” the restricted data to anyone not properly authorized. However, if a researcher has questions we are allowed to analyze the data and provide the results to the researcher and do not violate the restricted use clauses.

The advantage of having numerous restricted data sets is the ability to merge data from different data sets into one data set. The NCES data sets have common variables and the assignment of characters or numerics to represent these variables is consistent. Therefore, if I am working with the 1972 data set on high school performance, I can merge this data set with data from the 1988 National Educational Longitudinal Survey (NELS:88) data set. This will allow for direct comparisons of schools, districts, and states on educational performance to be conducted at regional and national levels. However, the procedures for merging data sets can be very difficult and confusing unless you have a solid understanding of data management. The use of SAS programming in developing this aspect was crucial to the goal of this project. The success of this process for merging the data sets also provides a wonderful new avenue for conducting research using the NCES data sets.

Data Acquisition

Data was requested from NCES for 12 public data sets and 3 restricted use data sets. The data sets included both cross-sectional and longitudinal data. The various data sets addressed different types of information on students. For example, the National Assessment of Educational Progress (NAEP) Data set examines educational performance of students and is collected every other year. The data collected for the NAEP data set is designed to allow for comprehensive review of the educational progress of students on various science, math and reading scales. However, the data is cross-sectional and does not allow for direct longitudinal comparisons. The National Education Longitudinal Study (NELS:88) data set is a longitudinal data set tracking students as they progress through high school and beyond. Although these two data sets have a different emphasis, they do contain variables in common such as school district, state, and demographics. Thus, these two data sets can be combined, merging on one of the common variables, to allow for certain types of aggregated information to be compared. There was an additional 10 data sets which were requested based on the same primary characteristic that they have variables in common which allows for the data sets to be merged. Additionally, there are any number of other data sets which could also be included in this database, such as local and state information, which could then be merged with some of the national information to allow for any number of larger comparisons. A key element to this type of data merging is identifying a common variable and making sure that the coding is exactly the same in both data sets. Once this is done, then any number of new types of subsets of data are possible and can be generated for any potential research project.

Sample of Programming

Key elements for merging and combining the data are the **Merge**, **Sort** and **Output** commands.

Example:

```
Data One; infile “d:cdrom1";
input x1 x2 x3 x4 x5 SSNUM;

Data Two; infile “c:\quant\data2";
input y1 y2 y3 y4 y5 SSNUM;

/* note the variable SSNUM is present in
both data sets */

Proc Sort Data=One; by SSNUM;
Proc Sort Data=Two; by SSNUM;

Data Three;
  Merge One (keep = x1 - x3 SSNUM
          in=one)
            Two (keep = y1 - y3 SSNUM
        in=two);
  By SSNUM;
run;

/* Next, an output data set is created so the
researcher can continue to utilize this data set
without having to sort and merge the data */

Data _Null_;
  Set Three;
  File “a:newdata”;

/* Next is an example of how to place the
data in a fixed column format */
```

```
  Put @1 (SSNUM) (8.) @9 (x1 - x3) (1.)
@12 (y1-y3) (1.);
run;
```

This simple program provides a model of the process necessary to merge and create a new data set. For example, in the preceding program data sets one and two are merged. The variable SSNUM is in both data sets and is necessary to merge the data. If there is no common variable the data can still be merged, but the data will be concatenated to the already existing data and not actually “merged.” In order to merge the data, both data sets must be sorted by the variables which you are using to merge the data.

Another aspect provided in this example is that the researcher does not intend to use all the data from both data sets. Instead, the researcher would like to use only a selected number of variables from each (x1 - x3 and y1 - y3). Using the KEEP command a subset from data sets one and two is used to create data set three.

A final model is provided to show how you can create an output data set for later use. In the example program an output data file called “newdata” is created and stored in the “a” drive. The data could be stored on any drive or in a directory on the “c” drive. This purpose here was to demonstrate that this process is easy and useful for researchers who do need all the data in several data sets, but only some of the variables.

Summary

Ultimately, the goal was to develop a procedure for prospective researchers to request data from the University of Arkansas Educational database. A preliminary/prototype SAS program has been developed to facilitate and operationalize all the goals of this project. The outcome has been the ability to use a primary program, written in SAS, to create subsets of data with the requested variables and all the necessary statistical issues addressed so that the researcher(s) could then complete any proposed research project requiring educational statistics or data.

Sean W. Mulvenon
241 Graduate Education Building
University of Arkansas
Fayetteville, AR 72701
Phone: (501) 575 - 8727
E-mail: seanm@comp.uark.edu