Data set Management Procedures for Developing a Prevention and Counseling Database.

Sean W. Mulvenon, University of Arkansas, Fayetteville, AR
Sherry Ceparich, Arizona State University, Tempe, AZ
Barbara Weber, Arizona State University, Tempe, AZ
Arlene Metha, Arizona State University, Tempe, AZ

Abstract

The use of the SAS statistical package has been very instrumental in the development of two separate databases at Arizona State University (A.S.U.). The databases are the result of studies which investigate the use of various interventions designed at reducing at-risk behavior in middle and high school students identified as potentially suicidal. The data for each subject was collected from numerous agencies and the problem of data management, i.e., the process of merging and organizing the data is the basis for this paper. The SAS packages utilized were BASE, SAS/STAT, and IL. This program is designed to work on any DOS operating system which can operate the SAS program and for users with intermediate expertise in using the SAS programming language.

Introduction

The use of large databases in educational and psychological research has grown in the last few years. The use of computerized testing and scoring has led to an increase in the number of instruments and surveys developed to evaluate educational progress and psychological behavior. The availability of these instruments and surveys has ultimately led to research of individuals on a much broader scope and the need to create and manage large databases which are very diverse in the types of information maintained. However, the issue of data management develops when a researcher attempts to combine differents types of information on individuals, maintained in different systems and formats, into one format to allow for a more thorough or comprehensive investigation of these individuals in prospective studies.

The use of packages such as Database IV and Excell and other spreadsheets is common because of their ease to explain and utilize. However, when managing numerous data sets, these packages are not as effective as SAS. Issues in data management such as data manipulation, merging data sets, creating output data sets of specific variables of interest, and data analysis are simply not avialable at the level necessary for most comprehensive studies. The purpose of this paper is to address the issues and procedures utilized in SAS to develop and maintain a database which would provide solutions to the previously mentioned issues when developing and maintaining a large database.

Method

The use of large databases and the management of these databases has become somewhat limited due to the plethora of ways in which data can be obtained. Further, the collection and subsequent development of this database needed to be completed in such a manner that the database could be made available for use by any prospective researcher. Things to consider included the naming of variables, how variables are named, how data sets are merged, collapsing across groups, the transferring of data from one form or system to the next the system. This paper address the process by which this has been accomplished for research completed at Arizona State University.

The data used for this project was collected measured different types of demographic variables and psychological constructs and was obtained from a number of sources and subsequently contributed to the need to develop a procedure for merging and compiling data.

Instruments

The Children's Depression Inventory (CDI). The (CDI), is a 27 item self-rating scale designed to assess symptoms of depression in children ages 8 to 17 years (Kovacs, 1992). The (CDI) provids a total measure of depression, which can be further partitioned into submeasures of negative mood, interpersonal problems, ineffectiveness, anhedonia, and negative self-esteem.

The hopelessness Scale for Children (HSC).

The (HSC) is a self report measure which assesses children's negative expectations toward the future. The (HSC) has seventeen dichotomously scored items and scores may range from 0 to 17.

Rosenberg Self-Esteem Scale. The Rosenberg Self-Esteem scale is a global measure of self-esteem with possible scores on 10 items of 1 to 4 and total scores ranging from 10 to 40 (Rosenberg, 1965).

Subtance Use Scale. Substance use was measured by an abbreviated version of the Arizona Department of Education Substance Use Scale. The 10-item scale assessed the individual's self-reported use of drugs and alcohol within the past year or within the last 30 days. Possible scores for this survey were from 1 to 5 and total scores of 10 to 50.

Suicide Risk Measure. Suicide risk was operationally defined as a spectrum of suicidal behaviors which included attitudes toward suicide, suicide ideation, suicide attempt, and exposure to suicidal behavior. For this study, four suicide risk domains were tapped by a self-report measure which included attitudes toward suicide, suicide ideation, suicide attempt, and exposure to suicidal behavior. The questions (8 items) were derived from existing instruments published in the suicidiology literature (i.e., Shaffer, Garland, Underwood, and Whittle, 1987). The scoring of these items was on an individual basis and could range from 0 to 5.

Coping Response Inventory. The (CRI) was designed to assess the coping responses of adolescents ages 12 - 18 and may be used with children who may be classified as healthy to those as having psychiatric, emotional, or behavioral problems (Moos, 1993). The (CRI) consists of 48 dichotomously scored questions and can be partitioned into 8 separate categories of coping. Further, these 8 categories can be partitioned into two general categories of avoidance and approach coping categories. Total scores may range from 0 - 48, scores on one of the 8 subscales may range from 0 to 8, and the scores on the avoidance/approach subscales may range from 0 to 24.

Stress Inventory. The Stress Inventory (SI) is designed to assess the type and degree of stress identified by an individual. The inventory consists of 22 items which have two components; a dichotomous (yes/no) response to determine if a situation caused positive or negative stress, proceded by a response which addresses the degree of stress caused by the situation. The scores for this survey are recoded with negative responses (no) coded as negative numbers, and all responese (yes/no) multiplied by the degree of stress. Thus, scores can potentially range from -88 to +88.

Data Management Issues

There were a number of specific issues which needed to be addressed in order to develop and manage this database. First, each of these data sets needed to be coded (machine scoring) and stored as ASCII text files with specific column specifications. Second, a common variable had to be identified in each data set to provide a means for merging the datasets so that no information would be lost or concatenated at the bottom of the data set (i.e., stacking of the various data sets on top of each other). Third, there were a number of data manipulations which would be necessary to correctly score the various inventories. Thus, the questions of when and where would be the most appropriate places to complete the data manipulations and further, how to complete these manipulations? Finally, developing a program that allows for all these issues to be addressed and provides the flexibility to create specific output data sets which could include any raw values, rescored values, or scores on any subset or totals of variables from the various data sets.

Result

A program was developed which we believe addresses the important issues already discussed. All the data was accessed and merged utilizing one program. Further, if data manipulations were necessary to rescore any values, the new values were provided with different names. Thus, all the the data, both raw and rescored, could be accessed and utilized from one program. The key elements of the SAS programming language utilized were the MERGE and OUT commands. A sample program is provided to demonstrate how this final program was developed.

Example

/* We began by identifying the data sets of interest and their location on the harddrive */

```
Data One; infile "c:suicide1.dat";
 input x1 x2 x3 x4 x5 IDNUM;
run;
```

```
        Data Two; infile “c:suicide2.dat";
        input y1 y2 y3 y4 y5 IDNUM;

/* Next is an example of necessary data
manipulations */

        ny1= 6 - y1;

/* This data manipulation will rescore data which is
originally a 5 to 1, 4 to 2, ...., 1 to 5 */

/* This process was repeated for all seven data sets */

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

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

        Data Three;
          Merge One (keep = x1 - x5 IDNUM
                        in=one)
                  Two (keep = y1 - y5 ny1 IDNUM
                              in=two);
          By IDNUM;
        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 - x5) (1.)
        @14 (y1-y5) (1.) @19 (ny1) (1.);
        run;
```

The key element in this program is the creation of a new data set which includes the variables from all the data sets (data set three in the example). The output data set , Data _Null_ can be modified to create new data sets which are just subsets of data three for researchers who are not interested in using all the data[1].

Summary

The purpose of this paper was to address the issue of merging data sets in social science research. The emergence of computerized testing and the creation of numerous instruments to measure various pyschological constructs has resulted in large amounts of data, but in different formats. The purpose of this papers was to demonstrate that data from different researchers, but from the same subjects, can be merged to create larger composites of each subject and greater potential for research. The data set created at Arizona State University has now been used for numerous articles and several dissertations because the merging of a number of data sets has created the potential to study a vast number of psychological issues.

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

[1]Any person who would like a copy of the actual program developed in this study can obtain one by requesting in writing a copy from Sean W. Mulvenon, 241 Graduate Education Building, University of Arkansas, Fayetteville, AR, 72701.