THE USE OF SAS FOR THE ANALYSIS OF FAMILY STUDIES
Deborah Powell, J. Philip Miller and C. Robert Cloninger
Washington University School of Medicine
St. Louis, Missouri 63110

ABSTRACT


The analysis of family studies present significant problems with respect to the use of standard statistical packages as the data set does not conform to the simple rectangular management facilities of SAS, however, provide the necessary tools to produce the desired analyses.

The principal challenge of family studies is the variety of ways in which the data needs to be organized for the production of the required analyses. Some analyses require only the selection of an appropriate subset of the database, e.g. all probands, all fathers, all first degree relatives of probands. Others require linking of related individuals, e.g. mother-daughter, father-offspring, husband-wives. Since the amount and/or type of information often depends upon the familial relationship of the individual, we have found that an effective strategy is to store the data in multiple SAS data sets according to the type of information obtained, but all within a single SAS library. Analysis data sets are then straightforwardly developed with the use of sorting and merging operations within SAS. These analysis data sets can either be created as temporary, work data sets or saved in the library if they are to be frequently utilized.

INTRODUCTION

The analysis of family studies can be difficult when the data set contains several classifications for family members. The number of different ways of looking at the data is almost unbounded. Imagine trying to analyze family data with all its different patterns using conventional statistical packages such as SPSS or BMDF. In the Department of Psychiatry at Washington University, we have found SAS to be a most valuable instrument in management and manipulation of family data. This paper will discuss some of the studies in which we have been using SAS quite successfully and indicate the approaches and techniques which we have found useful.

The first study is one on 80 fathers who were part of a previous study of 223 black men in St. Louis initially studied in 1965-66 (Robins, et. al. 1975). They were followed up in 1970-72. With this followup, information from school records and police records were collected on their wives and children.

Three SAS data sets were created. The father data set contained approximately 800 variables and 80 observations. The mother data set contained 90 variables and 80 observations (there are more mothers than fathers because 8 of the fathers had been married twice). The child data set contained 180 variables and 159 children each (there were 1 to 4 children born to the first marriages of the fathers and either 1 or 2 born to the second marriages).

The second was a family history study which served as a pilot study for an alcoholism and criminality study that is presently being conducted. Three-hundred forty felons were interviewed with questions asked of them about themselves and also questions were asked about each of their family members. Information was collected on grandparents, parents, siblings, half-sibs, spouses, children, etc. in terms of the felon's view of his family.

Through examples from these studies, others may see the importance of planning a strategy of analysis and be able to use some of these techniques for their own data analysis tasks. Family studies are by no means an easy thing to deal with but several SAS features greatly facilitate the task. We make heavy use of Renaming, Macro's, Selecting, Sorting and Merging of data sets.

The general strategy which we have found to be useful is to create multiple SAS data sets within a library, each data set being composed of those observations where we have comparable information on each subject. This usually provides more efficient storage than simply-forming the information into a single rectangular data set. Each data set also contains variables which enable that observation to be linked with others in the family through the use of sorting and merging. Generally we create the linked data set appropriate for a given analysis each time that an analysis run is made. Not only does this conserve on data storage space but it simplifies error correction procedures as each datum is stored in only one place.

This project has been supported by the National Institute of Mental Health, Grant Nos. MH 18864, AA 00209, AA 03539, and ME 00048.
in the first example we made use of the father, mother, and child data sets. This particular job (Figure 1) looked at children whose mothers dropped out of high school predictive of mate children. A1BO mother's truancy is related to having illegitimate children. A1BO mother's truancy by reflecting the total number of quarters after kindergarten in which she had poor attendance. It ranged in value from 0 to 99 with 0 representing no information and 99 for no quarters with poor attendance. Five or more quarters with poor attendance was defined as "truant" (M0417A=1) and fewer quarters as "non-truant". The variable M0442 was the year final high school drop out occurred for the mother. It ranges in value from 0 to 9 with 0 standing for no information and 9 for graduated (1 = 1st year dropped out, 2 = 2nd year dropped out, etc.). The years dropped out were then grouped to represent "did not graduate" (M0442A=1) and "graduated". Four child data sets were then created (Figure 3) using CHLDPOS which indicates the child's position in his particular family. For each child data set, the Macro REC (Figure 4) was used to create a number of composite variables. A1BO mother's truancy is related to having illegitimate children.

In the next example, we make use of the mother and child files to explore whether the mother's truancy is related to having illegitimate children. Also, was the fact that a mother dropped out of high school predictive of her marriage breaking up? First. (Figure 2) a temporary mother data set was created (MO) was created. Only the variables needed for the analysis were kept. This data set was sorted into order by the father's ID. Next, a temporary father-mother data set (FAMO) was created by merging with the father's ID (CPAID). From the father's data set (FAMILY.FATHER) only the variables needed for creating composite variables plus those to be used for the analysis in their original form were kept. In so doing both time and money were saved since only 13 of the 800 variables stored on the father were retained. This father-mother data set consisted of 98 observations (the variables for the father were repeated for the 8 additional mothers) representing the 88 marriages that produced children.

Next, a temporary child data set was created (CH) and sorted by the mother's ID (CMOID). The father-mother data set was resorted and merged with the child data set. It is important to have used the mother's ID for the merging since by using the father's ID, the second wife might have been merged with children from the first marriage.

In the next example, we make use of the mother and child files to explore whether the mother's truancy is related to having illegitimate children. Also, was the fact that a mother dropped out of high school predictive of her marriage breaking up? First. (Figure 2) a temporary mother data set was created (MO). On it was kept only the variables needed for the analysis. The variable M0417 defines mother's truancy by reflecting the total number of quarters after kindergarten in which she had poor attendance. It ranged in value from 0 to 99 with 0 representing no information and 99 for no quarters with poor attendance. Five or more quarters with poor attendance was defined as "truant" (M0417A=1) and fewer quarters as "non-truant". The variable M0442 was the year final high school drop out occurred for the mother. It ranges in value from 0 to 9 with 0 standing for no information and 9 for graduated (1 = 1st year dropped out, 2 = 2nd year dropped out, etc.). The years dropped out were then grouped to represent "did not graduate" (M0442A=1) and "graduated". Four child data sets were then created (Figure 3) using CHLDPOS which indicates the child's position in his particular family. For each child data set, the Macro REC (Figure 4) was used to create a number of composite variables. A1BO mother's truancy is related to having illegitimate children. Also, was the fact that a mother dropped out of high school predictive of her marriage breaking up? First.
that the family had been split up since the child's birth, but before the child went to school or split up during the child's school age years. The use of the Macro saved coding time in creating these regrouped variables for each child data set. Some variables were renamed so that after merging the mother with her children it is possible to see more clearly if any child was illegitimate or if any child came from a broken home.

FAMILY HISTORY STUDY

Turning to the felon study, the example comes from a more basic starting point in the handling of the massive bulk of data collected. As part of the data cleaning activity it was desirable to be able to look at an entire family in order to see if there were inconsistencies in the information between different family members. For example, children with different fathers may have been erroneously coded as full sibs. In addition, examining complete families one by one might give the researcher a greater feeling for the data, and consequently, hypotheses might be developed and examined by use of techniques similar to those outlined above. The pedigree-like format for the information was also needed in order to be able to interface with other specialized programs for quantitative genetic analyses.

To create the output, a data set (GENERATE) was developed that contained a table of the relations possible in the pedigree (Figure 5).

*FIGURE 5:
DATA GENERATE;
LABEL GEN=GENERATION NO;
REL=SEX OF RELATION TYPE;
CARDS:
1 0 .
2 -1 F
3 -1 M
4 0 M
5 0 F
. .
78 0 F

This table described the relationship between the proband and each of his family members. RELATE was a code for the relationship to the proband i.e. the proband himself/herself, father, mother, etc. (so that 1 = proband, 2 = mother, 3 = father, 4 = brother, 5 = sister, etc.). GEN indicated a number as to which generation (relative to the proband) that particular family member belonged (e.g. -1 = a generation before the proband, such as parents, aunts, uncles; 0 = the proband's own generation, such as brothers, sisters, spouse; +1 = the generation after the proband, such as his/her offspring). REL=SEX assigned the proper sex code to each family member. Next, (Figure 6) it was necessary to create a temporary file containing all the family members, probands included, and keeping pertinent information about each one -- such as

+FIGURE 6;
DATA DECK1;
SET HALFSIB.DECK2 (KEEP=RELATION FAMILYID
MEM NUM INDIV_ID SEX RACE AGE BIO_MO
BIO_FA LIVEMENT FIRSTAGE LAST_AGID)
RELATE=REL(PRELATION/100)
SPOUSE=MOD(RELATION,10)=0;
FORMAT RELATION BIO_MO BIO_FA 24;
PROC SORT DATA=DECK2; BY INDIV_ID;
PROC SORT DATA=HALFSIB.DECK5 (KEEP=INDIV_ID DX)
OUT=DECKS; BY INDIV_ID;
DATA FOLKS; MERGE DECK2 DECKS; BY INDIV_ID;

the relation to the proband, sex, race, age, biological mother, biological father, whether the family member ever lived with the proband or not, and if so, first and last age lived with him/her. The relationship of the individual to the proband was decoded from the first two digits of a four digit RELATION variable. Sorting was done using the individual's unique ID. Another temporary data set (DECKS) was created retaining on it, the individual's ID and his/her diagnosis (sociopathy, hysteria, alcoholism, drug dependence, sexual deviation, etc.). These two data sets were then merged using the individual's ID.

The resultant data set, FOLKS, was sorted by RELATE (Figure 7) so now it was in such an order that all the probands were first, followed by the mothers, then the fathers, etc. FOLKS with its 4776 observations was merged with GENERATE, the data set created in the very beginning of this job. The subsetting IF statement was used to eliminate any of the 78 possible relations that didn't actually exist within this data set. For instance, a code provided for the possibility that a proband's son might have a daughter is possible but if no son actually did, this code would not appear. FOLKS was sorted in order by the family ID, generation, and relation. For each family, then, information about each member was in order by generation, and within generation by the relation to the proband.

For the sake for consistency checks, information asked of the proband concerning number of brothers, aunts, half-brothers, etc. was pulled from the proband's interview (HALFSIB.DECK1), sorted in order by FAMILYID, and merged with FOLKS by FAMILYID.

Utilizing the printing capabilities of SAS, a family pedigree was then produced for each felon (Figure 8). Figure 9 shows how the information was organized within each cluster for each family member. The SAS statements for
**FIGURE 8:**

FAMILYID=12 PRO_RACE=1 PROBSEX=1 PROB_AGE=24 BROTHERS=4 SISTERS=1 HALF_BRO=0 HALF_SIS=0 STEP_BRO=0 STEP_SIS=0 POST_FA=0 POST_SON=0 daughteR=1 POST_DAD=0 OTHER_BIO=4 OTH_UNRE=3 TOTAL=16

**GENERATION = -2**

<table>
<thead>
<tr>
<th>maternal</th>
<th>paternal</th>
<th>maternal</th>
<th>paternal</th>
</tr>
</thead>
<tbody>
<tr>
<td>grandfather</td>
<td>grandmother</td>
<td>grandfather</td>
<td>grandmother</td>
</tr>
<tr>
<td>13/158</td>
<td>12/157</td>
<td>14/159</td>
<td>15/160</td>
</tr>
<tr>
<td>1200 1 1 C</td>
<td>1300 2 1 C</td>
<td>1400 2 1 C</td>
<td>1500 1 1 C</td>
</tr>
<tr>
<td>1700 1600</td>
<td>1900 1600</td>
<td>2300 2200</td>
<td>2100 2000</td>
</tr>
<tr>
<td>NO 20</td>
<td>NO 20</td>
<td>NO 20</td>
<td>NO 20</td>
</tr>
</tbody>
</table>

**GENERATION = -1**

<table>
<thead>
<tr>
<th>other</th>
<th>father</th>
</tr>
</thead>
<tbody>
<tr>
<td>2/147</td>
<td>1/146</td>
</tr>
<tr>
<td>0200 2 1 42</td>
<td>0300 1 1 52</td>
</tr>
<tr>
<td>1300 1200</td>
<td>1400 1500</td>
</tr>
<tr>
<td>2 18 42 . 20</td>
<td>2 28 36 10</td>
</tr>
</tbody>
</table>

**GENERATION = 0**

<table>
<thead>
<tr>
<th>proband</th>
<th>proband's first spouse</th>
<th>proband's second spouse</th>
<th>brother 1</th>
<th>brother 2</th>
<th>brother 3</th>
<th>brother 4</th>
<th>sister 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0100 1 1 24</td>
<td>0101 2 1 25</td>
<td>0102 2 1 18</td>
<td>0103 1 1 22</td>
<td>0200 1 1 20</td>
<td>0300 1 1 20</td>
<td>0400 1 1 20</td>
<td>0500 1 1 20</td>
</tr>
<tr>
<td>0200 0300</td>
<td>0910 5810</td>
<td>0920 5820</td>
<td>0200 0300</td>
<td>0200 0300</td>
<td>0200 0300</td>
<td>0200 0300</td>
<td>0200 0300</td>
</tr>
<tr>
<td>4 4 23 24 20</td>
<td>5 18 18 4 3 3 20 20</td>
<td>5 3 21 8</td>
<td>3 3 20 20</td>
<td>5 18 18</td>
<td>4 5 18 18</td>
<td>4 NO 19</td>
<td></td>
</tr>
</tbody>
</table>

**GENERATION = 1**

<table>
<thead>
<tr>
<th>daughter</th>
<th>step-daughter</th>
</tr>
</thead>
<tbody>
<tr>
<td>8/153</td>
<td>11/156</td>
</tr>
<tr>
<td>0101 2 1 6</td>
<td>5102 2 1 3</td>
</tr>
<tr>
<td>0101 0300</td>
<td>0102 5830</td>
</tr>
<tr>
<td>4 1 1 20</td>
<td>4 3 3 20</td>
</tr>
</tbody>
</table>

**NUMBER OF FAMILY MEMBERS=16**

---

**FIGURE 9:**

PROIDEN ID NUMBER INDIVIDUAL ID NUMBER
RELATION SEX RACE AGE
BIO MD 310 FA
LIVED WITH PROBAND FIRSTAGE LASTAGE DX

Producing this output are shown in Figure 10. The information for family 12 which is listed at the top of Figure 8 comes from HALFSIB.DMK1. The proband was a white male, age 24, who came from a family of 16 members, including himself. He had four brothers, one sister, and one biological daughter. His ex-wife, present wife and stepdaughter were grouped in OTH_UNRE and his grandchildren were grouped in OTH_BIO.

Below this family description were printed clusters of data describing each family, member, facilitating the checking of the correctness of the coding. Each member's sex agreed with the relationship code (e.g., all brothers were male, all sisters were female, the race of the parents and children agreed, the entire family was white). The proband was uncertain about his grandparents ages but knew the rest of the members ages. His mother was 42, his father, 52. His first wife and sister were older than he and his second wife and four brothers, younger. His daughter was six and his step-daughter, three. The proband, along with all his brothers and sister had the same parents. Each member who lived with the proband was coded properly for LIVWITH, and in observing the ages you can see that the proband lived with the father only nine years whereas he lived with his mother for twenty-five years. Another interesting point that might require checking, is the fact his sister never lived with him. The proband's biological daughter lived with him through age one and his step-daughter and second wife were currently living with him.

As far as diagnoses were concerned, the proband's grandparents, mother, first spouse, second brother, and two daughters, had no psychiatric illness. His father had questionable alcoholism, his first brother had severe alcoholism, and his sister had some psychiatric illness not defined. The proband, along with his second wife and third and fourth brothers all had questionable sociopathy.

From this one family the benefits of creating the output are apparent. Not only is the investigator able to question the validity of the data from the proband's record and each family member's record, but also is able to see...
The itemized version can be administered only by a psychiatrist while the verbatim version can be administered by a psychiatrist or a research assistant. Every patient is given one interview using the itemized version. The other two interviews can be any combination of psychiatrist verbatim and research assistant verbatim.

Unfortunately, it was not always possible to obtain all three interviews. Sometimes patients became too sick to interview, sometimes they checked out of the hospital, and sometimes the patient refused after sitting through two interviews. The patterns that can be tested for reliability were AB, AC, BC, BB, and CC (A = Psychiatrist Itemized, B = Psychiatrist Verbatim, and C = Research Assistant Verbatim). In the last example, pairs who disagreed on Anxiety Neurosis using the AC pattern were examined. We wanted to see on what symptoms the pairs disagreed so that the discrepancies on the diagnosis could be understood better.

First, two temporary files were created—one (PSYCHI) containing the Psychiatrist Itemized (A) and the other (RAV) containing the Research Assistant Verbatim (C) (Figure 11).

```
+FIGURE 10 ;
DATA NULL ;
FILE PRINT LINE=1;
SET PAGES ; BY FAMILYID GEN ;
IF FIRST.FAMILYID THEN LINK NEW_FAM ;
IF FIRST.GEN THEN LINK NEW_GEN ;
#1 ; *COUNT NUMBER OF FAMILY MEMBERS ;
PUT #1 $COL NR NUN INDIV_ID ;
L1=LINE+1 ;
PUT #1 $COL RELATION SEX RACE AGE ;
L1=LINE+1 ;
PUT #1 $COL BIO MO MO_F ;
L1=LINE+1 ;
IF LIVETHEN = 1 ;
THEN PUT #1 $COL RELATION FIRSTAGE LASTAGRN ;
IF LIVETHEN = 1 ;
THEN PUT #1 $COL 'NO' ;
IF COL > 100 THEN LINK NEW_LINE ;
RETURN ;
NEW LINE ; IF L >= 55 THEN PUT PAGE FAMILY ID = ' (CONTINUED )' ;
LINE = L ;
RETURN ;
NEW GEN ; IF N > 1 THEN PUT // 'NUMBER OF FAMILY MEMBERS = N ' ;
END RUN ;
```

The itemized version had its needed variables renamed. These were the symptom questions pertaining to Anxiety Neurosis. Notice the variable, INTROLE. This was the key to separating the different interviewers. Originally, it had three values, 1 = Psychiatrist Itemized, 2 = Psychiatrist Verbatim, and 3 = Research Assistant Verbatim. In this particular case to create the above two data sets was an easy task, though the comparison of a BB or CC pattern was difficult. INTROLE was redefined so that it served its purpose. The value 1 remained as equalling Psychiatrist Itemized, 2, 3, and 4 took on the values of only one Psychiatrist Verbatim, first of two Psychiatrist Verbatims, and second of two Psychiatrist Verbatims. Five, 6, and 7 took on the values of only one Research Assistant Verbatim, first of two Research Assistant Verbatims, and second of two Research Assistant Verbatims. Merging the two files (Figure 11) by ID and selecting only those cases for which there were observations in both data sets was the next step. Those pairs who disagreed on the diagnosis could be quickly identified and the symptoms for the Psychiatrist Itemized listed above the symptoms for the Research Assistant Verbatim.

The second data set had its needed variables renamed. These were the symptom questions pertaining to Anxiety Neurosis. Notice the variable, INTROLE. This was the key to separating the different interviewers. Originally, it had three values, 1 = Psychiatrist Itemized, 2 = Psychiatrist Verbatim, and 3 = Research Assistant Verbatim. In this particular case to create the above two data sets was an easy task, though the comparison of a BB or CC pattern was difficult. INTROLE was redefined so that it served its purpose. The value 1 remained as equalling Psychiatrist Itemized, 2, 3, and 4 took on the values of only one Psychiatrist Verbatim, first of two Psychiatrist Verbatims, and second of two Psychiatrist Verbatims. Five, 6, and 7 took on the values of only one Research Assistant Verbatim, first of two Research Assistant Verbatims, and second of two Research Assistant Verbatims. Merging the two files (Figure 11) by ID and selecting only those cases for which there were observations in both data sets was the next step. Those pairs who disagreed on the diagnosis could be quickly identified and the symptoms for the Psychiatrist Itemized listed above the symptoms for the Research Assistant Verbatim.

Besides the reliability study and the family patterns within a family and to formulate thoughts for hypotheses to be tested. Does alcoholism run in families? Is it associated with certain other forms of psychopathology such as sociopathy, depression? What effects are felt by the proband when he has to live with an alcoholic father? Questions like these are only the beginning.

**DIAGNOSTIC INTERVIEW RELIABILITY STUDY**

Finally, it is important to indicate that the techniques and applications used in these examples for family studies can be used for other studies. Currently at Renard Hospital we are administering a generalized interview to 120 patients covering specific psychiatric diagnoses (i.e. Depression, Hysteria, Schizophrenia, Alcoholism, Anxiety Neurosis, etc.). This interview is being administered to a patient on three successive days by three different people. The purpose is to test the reliability of a new verbatim version of the interview as compared with the standard itemized version.

The itemized version can be administered only by a psychiatrist while the verbatim version can be administered by a psychiatrist or a research assistant. Every patient is given one interview using the itemized version. The other two interviews can be any combination of psychiatrist verbatim and research assistant verbatim.

Unfortunately, it was not always possible to obtain all three interviews. Sometimes patients became too sick to interview, sometimes they checked out of the hospital, and sometimes the patient refused after sitting through two interviews. The patterns that can be tested for reliability were AB, AC, BC, BB, and CC (A = Psychiatrist Itemized, B = Psychiatrist Verbatim, and C = Research Assistant Verbatim). In the last example, pairs who disagreed on Anxiety Neurosis using the AC pattern were examined. We wanted to see on what symptoms the pairs disagreed so that the discrepancies on the diagnosis could be understood better.

First, two temporary files were created—one (PSYCHI) containing the Psychiatrist Itemized (A) and the other (RAV) containing the Research Assistant Verbatim (C) (Figure 11).
FIGURE 12;
DATA COMB;
MERGE PSYCH(IN=YES1) RAW(IN=YES2);
BY ID;
IF YES1 & YES2;
DATA DISAGREE;
SET COMB;
TITLE DISAGREE ON ANXIETY NEUROSIS;
IF RD0720 = RD0720A;
PUT ID 7. (RD0215-RD0221 RD0226) (3.)
(RD0227 RD0229) (5.) RD0255 3.;

studies presented, followup studies are an
excellent candidate for implementing the multi­
ple comparison techniques of SAS. Perhaps the
researcher would like to see what changes
occurred from time period one to time period two
and investigate possible reasons. Then again,
his/her might want to investigate information
collected only in the first interview or only
in the second interview. Whatever the reason,
it is not mandatory to create single rectangular
data sets that sometimes waste space and prove
inefficient.

All in all, the difficult task of handling
different types of data can be dealt with
rather efficiently using SAS. If you can out­
line what you want to do with your data, you
can then decide what type of sorting and merging
variables will be needed to obtain the data
sets desired and what type of selecting vari­
ables are needed to gather the proper cases into
the different data sets created for a given
analysis. Utilizing the RENAME and MACRO capa­
bilities along with the printing features in
SAS, the researcher has available to him/her the
procedures and statistics that will make his/her
task a feasible one.

ACKNOWLEDGEMENT

The authors would like to thank Dr. Lee
Robine for her advice, interest and constructive
criticism.

REFERENCE

Robins, L. N., Neut, P. A., Herjanic, B.;
Arrests and delinquency in two generations:
A study of black urban families and their
children, J. Child. Psychol. Psychiatry,