

PO01

A SAS® Macro: Similar Code for the Similar Schools Data Processing

Imelda C. Go, South Carolina Department of Education, Columbia, SC

ABSTRACT

In an effort to simplify and make more efficient the similar schools data processing for the South Carolina district and school report cards, a macro was written and documented for use among different programmers. The documentation provided clear instructions about how a programmer could provide parameters to the macro and obtain similar schools data that was normally provided by another programmer. Since similar schools data processing covered different types of data (i.e., primary school, elementary/middle school, high school, and district) and used the same algorithm but with different input variables, the situation lent itself well to macro programming. Given different variables (e.g. attendance rate, retention rate, etc.), the task was to find the median value for schools that are similar to a given school. Schools that are similar to a given school are those that have a poverty index that is within 5% of the given school's poverty index. For example, if a school's poverty index was 11.8%, then schools that are considered similar to it have a poverty index ranging from 6.8% to 16.8%. Once groups of similar schools have been identified, it becomes a simple matter to obtain the desired statistic for each group of schools.

For the purpose of providing an example for this paper, consider the following input SAS data set, named SchoolData, with variable names and values. There are three school types denoted by E, M, and H values for variable SchoolType.

ID	School	SchoolType	PovertyIndex	AttendanceRate	RetentionRate
01	Alpha Elementary	E	87.5	91.9	10.6
02	Beta Elementary	E	92.7	94.2	7.0
03	Kappa Middle	M	89.3	96.1	5.8
04	Gamma Middle	M	94.4	92.7	3.9
05	Kappa High	H	12.8	96.0	7.2
06	Epsilon High	H	29.8	94.4	7.3
07	Zeta School	E	41.5	97.6	10.5
08	Eta Academy	E	68.5	92.2	1.8
09	Theta Middle	M	79.1	90.9	3.2
10	Iota Middle	M	83.0	94.0	8.4
11	Psi High	H	28.3	97.2	0.8
12	Rho High	H	49.1	92.8	3.2
13	Mu Elementary	E	80.4	89.4	2.7
14	Nu School	E	69.4	93.1	8.7
15	Xi Elementary	E	97.6	94.3	6.7
16	Omicron Middle	M	74.2	96.1	7.3
17	Lambda High	H	73.3	94.1	2.8
18	Delta High	H	78.0	89.3	5.1
19	Sigma Elementary	E	60.6	89.5	3.8
20	Tau Elementary	E	18.8	98.2	0.5
21	Upsilon Middle	M	74.7	96.8	1.4
22	Phi Middle	M	23.2	91.6	7.4
23	Pi High	H	82.0	94.9	1.7
24	Chi High	H	83.0	92.8	7.2
25	Omega Elementary	E	24.5	97.3	10.2

GROUPS OF SIMILAR HIGH SCHOOLS

High schools are identified when SchoolType is H.

ID	School	SchoolType	PovertyIndex	AttendanceRate	RetentionRate
05	Kappa High	H	12.8	96.0	7.2
06	Epsilon High	H	29.8	94.4	7.3
11	Psi High	H	28.3	97.2	0.8
12	Rho High	H	49.1	92.8	3.2
17	Lambda High	H	73.3	94.1	2.8
18	Delta High	H	78.0	89.3	5.1
23	Pi High	H	82.0	94.9	1.7
24	Chi High	H	83.0	92.8	7.2

If we sort the schools according to increasing PovertyIndex values, we get the following table.

ID	School	SchoolType	PovertyIndex	AttendanceRate	RetentionRate
05	Kappa High	H	12.8	96.0	7.2
11	Psi High	H	28.3	97.2	0.8
06	Epsilon High	H	29.8	94.4	7.3
12	Rho High	H	49.1	92.8	3.2
17	Lambda High	H	73.3	94.1	2.8
18	Delta High	H	78.0	89.3	5.1
23	Pi High	H	82.0	94.9	1.7
24	Chi High	H	83.0	92.8	7.2

The following table summarizes the different sets of similar schools (denoted by 1 through 8). Each school is associated with a group based on a range that is $\pm 5\%$ of a particular school's PovertyIndex value. Two high schools are in the same group if they have a group number in common based on the cross-tabulation below. It is possible that a school is similar to no other school (e.g., Kappa and Rho Highs). Different schools might be associated with exactly the same groups (e.g., groups 2 and 3 are identical). A school could be a member of several groups (e.g., Delta High is in groups 5, 6, 7, and 8).

Poverty Index	Poverty Index Range for Group	Group	ID	School	Kappa High	Epsilon High	Psi High	Rho High	Lambda High	Delta High	Pi High	Chi High
12.8	7.8 – 17.8	1	05	Kappa High	1							
28.3	23.3 – 33.3	2	06	Epsilon High		2, 3	2, 3					
29.8	24.8 – 34.8	3	11	Psi High		2, 3	2, 3					
49.1	44.1 – 54.1	4	12	Rho High				4				
73.3	68.3 – 78.3	5	17	Lambda High					5, 6	5, 6	6	6
78.0	73.0 – 83.0	6	18	Delta High					5, 6	5, 6, 7, 8	6, 7, 8	6, 7, 8
82.0	77.0 – 87.0	7	23	Pi High					6	6, 7, 8	6, 7, 8	6, 7, 8
83.0	78.0 – 88.0	8	24	Chi High					6	6, 7, 8	6, 7, 8	6, 7, 8

The desired similar schools statistics for a particular school will be derived using all other schools similar to the particular school. Hence, if a school is only similar to itself, then there will be no statistics for that group since there are no other similar schools to the particular school.

WHAT TO DO FOR ONE HIGH SCHOOL: EPSILON HIGH

One cycle of similar schools data processing for Epsilon High can be coded as follows on the left with comments on the right.

proc means data=SchoolData (where=(SchoolType='H' and round(28.3-5, .1)<= round(PovertyIndex, .1)<= round(28.3+5) and ID ne '06'))	Use the data from SchoolData, but restrict it to high schools (i.e., SchoolType='H'). Perform the calculations only for schools with a poverty index that is $\pm 5\%$ of 12.8%, which is the poverty index for Epsilon High. Exclude Epsilon High from the processing.
n median min max noprint maxdec=1; var AttendanceRate RetentionRate PovertyIndex; output out=EpsilonHigh median(AttendanceRate RetentionRate PovertyIndex) = MedianAttendanceRate MedianRetentionRate MedianPovertyIndex n(PovertyIndex) = NPovertyIndex min(PovertyIndex) = MinPovertyIndex max(PovertyIndex) = MaxPovertyIndex;	Obtain the N, MEDIAN, MIN, and MAX statistics. Do not print output for the procedure. Decimal values go to the tenths place only. Use the three variables listed in the VAR statement. Create an output data set called EpsilonHigh. Obtain the median for the three variables listed in parentheses and name them as indicated. Obtain the number, minimum, and maximum of poverty Indexes and name it as indicated.

The output data set EpsilonHigh will contain the following variables and values with PROC PRINT.

PROC PRINT Results for EpsilonHigh								
Obs	_TYPE_	_FREQ_	Median Attendance Rate	Median Retention Rate	Median Poverty Index	NPoverty Index	Min Poverty Index	Max Poverty Index
1	0	1	97.2	0.8	28.3	1	28.3	28.3

That is one down and many more to go. The task is to apply the same procedure to each of the high schools and to eventually produce one data set with all of the desired statistics from each high school. In the example for Epsilon High, we do not know for which school the statistics were generated for except through the name of the data set (i.e., EpsilonHigh). We also manually specified Epsilon High's poverty index of 28.3% in the WHERE= option. We will need several things in order to convert this lone example to a macro that will accomplish the task in an automated fashion for each of the schools.

There are certainly many different ways the macro can be developed. What follows is but one way to accomplish the task.

A MACRO VARIABLE WITH THE NUMBER OF OBSERVATIONS IN A DATA SET

We will use the following information from SAS technical support for the final macro.

Technical FAQ (523) Is there a way to determine the number of observations in a data set and ha - Microso...

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Refresh Print Mail New Window

Address <http://support.sas.com/faq/005/FAQ00523.html> Go Links Settings

Google G Go Bookmarks 0 blocked

support.sas.com > Documentation -- Navigate our Site --

Documentation

Home | Products & Solutions | Install Center | Operating Environments | FAQs

FAQ # 523

Q: Is there a way to determine the number of observations in a data set and have that value placed in a macro variable?

A: Yes, by using the following macro entitled NUMOBS:

```
%macro numobs(dsn);
%global num;
data _null_;
if 0 then set &dsn nobs=count;
call symput('num', left(put(count,8.)));
stop;
run;
%mend numobs;

%numobs(dataset_name)
```

This will create a macro variable NUM that holds the number of observations in the data set.

If running Release 6.12 and above, you can use the new funtion %SYSFUNC along with the new ATTRN function, as follows:

```
%let dsid=%sysfunc(open(dataset_name));
%let num=%sysfunc(attrn(&dsid,nobs));
%let rc=%sysfunc(close(&dsid));
```

The first %LET statement opens the data set to be read. The second %LET assigns the macro variable NUM the number of observations in the data set. The third %LET statement closes the data set.

Product or Solution:
Base SAS

Operating System: All

Components:
BASE:Macro Facility

[Search](#) | [Contact Us](#) | [Terms of Use & Legal Information](#) | [Privacy Statement](#)
Copyright © 2005 SAS Institute Inc. All Rights Reserved.

Done Internet

TRANSITION TO MACROS

Let us produce the similar schools statistics for each high school in data set Subset.

```
data Subset;  
  set SchoolData;  
  if SchoolType='H';
```

Let us then put the number of observations in the data set into macro variable number using the following code.

```
%let dsid=%sysfunc(open(Subset));  
%let number=%sysfunc(attrn(&dsid,nobs));  
%let rc=%sysfunc(close(&dsid));
```

We will then iterate through the observations in HighSchools and for every high school in the data set, we will obtain the desired statistics using PROC MEANS.

```
%macro similar (SchoolType,BaseDsn);  
data subset;  
  set SchoolData;  
  if SchoolType="&schooltype";  
  
%do index=1 %to &number;  
%do;  
data _null_;  
  set subset;  
  if _n_=&index;  
  call symput ('id',id);  
  call symput ('school',school);  
  call symput ('poverty',PovertyIndex);  
  
data group; length groupschool $30.;  
retain groupid "&id" groupschool "&school";  
set subset;  
if round(&poverty-5,.1)<=  
round(PovertyIndex,.1)<=  
round(&poverty+5,.1) and id ne "&id";  
  
proc means data=group  
  n median min max  
  noprint  
  maxdec=1;  
  id groupid groupschool;  
  var AttendanceRate RetentionRate  
    PovertyIndex;  
  output out=StatsForOneSchool  
    median(AttendanceRate  
      RetentionRate Povertyindex) =  
    MedianAttendanceRate  
    MedianRetentionRate  
    MedianPovertyIndex  
  n(PovertyIndex) = NPovertyIndex  
  min(PovertyIndex) = MinPovertyIndex  
  max(PovertyIndex) = MaxPovertyIndex  
  format MedianAttendanceRate  
    MedianRetentionRate MedianPovertyIndex  
    5.1;  
  
proc append base=&basedsn  
data=StatsForOneSchool;  
%end;  
%mend similar;
```

Macro parameter SchoolType is for the school type and BaseDsn is for the name of the final data set. Data set subset contains just one type of school.

Iterate (1 till value in macro variable number) through all the observations in data set subset.
This data step is for the sole purpose of putting one school's id, school name, and poverty index in macro variables. Keep the ith observation (i is equal to the macro variable index value) in the temporary data set, which is taken from the list of schools in subset. Copy the values indicated into macro variables via SYMPUT.

For each observation in subset, assign fixed values to variables groupid and groupschool by using the retain statement. The group formed based on a particular school's poverty index will exclude the particular school itself. Only schools within 5% of the particular school's poverty index are included in the data set.

Now that the school ID and name around which the group was formed is on the data set, the values will be considered as ID values to appear in the output data set.

Once the statistics are produced, start compiling, via PROC APPEND, a data set whose name is specified in macro variable basedsn. The final data set in the PROC APPEND DATA= option will consist of all StatsForOneSchool records from each iteration.

Since we generalized the macro to accommodate different school types (E, M, H), we can now easily process for the three conditions by invoking the SAS macro thrice with different macro parameters.

```
%similar(H, StatsForHighSchools);
%similar(E, StatsForElemSchools);
%similar(M, StatsForMiddleSchools);
```

For SchoolType=H, the high schools will be processed according to their order in the original data set.

ID	School	SchoolType
05	Kappa High	H
06	Epsilon High	H
11	Psi High	H
12	Rho High	H
17	Lambda High	H
18	Delta High	H
23	Pi High	H
24	Chi High	H

DISCUSSION OF RESULTS

Iteration 1 (&index=1): For Kappa High, the macro routine will produce no observations because Kappa High is only similar to itself. Since the processing excludes the particular school upon which the school is based on, there are no results.

Data set group will be empty because of this.

```
NOTE: There were 8 observations read from the data set WORK.SUBSET.
NOTE: The data set WORK.GROUP has 0 observations and 8 variables.
```

Because data set group is empty, PROC MEANS will have no input and produce no output.

```
NOTE: No observations in data set WORK.GROUP.
NOTE: The data set WORK.STATSFORONESCHOOL has 0 observations and 10 variables.
```

Data set StatsForHighSchools will be empty because there are no observations for data set StatsForOneSchool.

```
NOTE: Appending WORK.STATSFORONESCHOOL to WORK.STATSFORHIGHSCHOOLS.
NOTE: BASE data set does not exist. DATA file is being copied to BASE file.
NOTE: There were 0 observations read from the data set WORK.STATSFORONESCHOOL.
NOTE: The data set WORK.STATSFORHIGHSCHOOLS has 0 observations and 10 variables.
```

Iteration 2 (&index=2): For Epsilon High, the macro routine will produce results because Epsilon High is similar to one school, Psi High. Data set group will contain all the schools that are similar to Epsilon High. Variables groupschool and groupid are for Epsilon High's identifying information.

```
NOTE: There were 8 observations read from the data set WORK.SUBSET.
NOTE: The data set WORK.GROUP has 1 observations and 8 variables.
```

Data Set GROUP after Epsilon High Iteration: PROC PRINT Results

2

Obs	groupschool	groupid	ID	School	School Type	Poverty Index	Attendance Rate	Retention Rate
1	Epsilon High	06	11	Psi High	H	28.3	97.2	0.8

Data set StatsForOneSchool will always have one observation, which contains the desired statistics.

```
NOTE: There were 1 observations read from the data set WORK.GROUP.
NOTE: The data set WORK.STATSFORONESCHOOL has 1 observations and 10 variables.
```

Data Set StatsForOneSchool after Epsilon High Iteration: PROC PRINT Results

Obs	groupid	groupschool	_TYPE_	_FREQ_	Median Attendance Rate	Median Retention Rate	Median Poverty Index	NPoverty Index	Min Poverty Index	Max Poverty Index
1	06	Epsilon High	0	1	97.2	0.8	28.3	1	28.3	28.3

Data set StastForHighSchools will receive its first observation from this iteration.

NOTE: Appending WORK.STATSFORONESCHOOL to WORK.STATSFORHIGHSCHOOLS.

NOTE: There were 1 observations read from the data set WORK.STATSFORONESCHOOL.

NOTE: 1 observations added.

NOTE: The data set WORK.STATSFORHIGHSCHOOLS has 1 observations and 10 variables.

Data Set StatsForHighSchools after Epsilon High Iteration: PROC PRINT Results

Obs	groupid	groupschool	_TYPE_	_FREQ_	Median Attendance Rate	Median Retention Rate	Median Poverty Index	NPoverty Index	Min Poverty Index	Max Poverty Index
1	06	Epsilon High	0	1	97.2	0.8	28.3	1	28.3	28.3

Iteration 3 (&index=3): For Psi High, only Epsilon High is similar to it.

Data Set GROUP after Psi High Iteration: PROC PRINT Results

Obs	groupschool	groupid	ID	School	School Type	Poverty Index	Attendance Rate	Retention Rate
1	Psi High	11	06	Epsilon High	H	29.8	94.4	7.3

We have, as expected, the desired statistics for Psi High.

Data Set StatsForOneSchool after Psi High Iteration: PROC PRINT Results

Obs	groupid	groupschool	_TYPE_	_FREQ_	Median Attendance Rate	Median Retention Rate	Median Poverty Index	NPoverty Index	Min Poverty Index	Max Poverty Index
1	11	Psi High	0	1	94.4	7.3	29.8	1	29.8	29.8

Data set StatsForOneSchool now receives its second record from the previous data set.

Data Set StatsForOneSchool after Psi High Iteration: PROC PRINT Results

Obs	groupid	groupschool	_TYPE_	_FREQ_	Median Attendance Rate	Median Retention Rate	Median Poverty Index	NPoverty Index	Min Poverty Index	Max Poverty Index
1	06	Epsilon High	0	1	97.2	0.8	28.3	1	28.3	28.3
2	11	Psi High	0	1	94.4	7.3	29.8	1	29.8	29.8

Iteration 4 (&index=4): For Rho High, no observations result because no other school is similar to it just as Kappa High had no other similar schools.

Iteration 5 (&index=5): For Lambda High, only Delta High is similar to it. Data set StatsForOneSchool now has three records.

Data Set StatsForHighSchools after Lambda High Iteration: PROC PRINT Results

Obs	groupid	groupschool	_TYPE_	_FREQ_	Median Attendance Rate	Median Retention Rate	Median Poverty Index	NPoverty Index	Min Poverty Index	Max Poverty Index
1	06	Epsilon High	0	1	97.2	0.8	28.3	1	28.3	28.3
2	11	Psi High	0	1	94.4	7.3	29.8	1	29.8	29.8
3	17	Lambda High	0	1	89.3	5.1	78.0	1	78.0	78.0

Iteration 6 (&index=6): For Delta High, three other schools are similar to it.

Data Set GROUP after Delta High Iteration: PROC PRINT Results

Obs	groupschool	groupid	ID	School	School Type	Poverty Index	Attendance Rate	Retention Rate
1	Delta High	18	17	Lambda High	H	73.3	94.1	2.8
2	Delta High	18	23	Pi High	H	82.0	94.9	1.7
3	Delta High	18	24	Chi High	H	83.0	92.8	7.2

Data set StatsForOneSchool now has a total of four records.

Data Set StatsForOneSchool after Delta High Iteration: PROC PRINT Results

Obs	groupid	groupschool	_TYPE_	_FREQ_	Median Attendance Rate	Median Retention Rate	Median Poverty Index	NPoverty Index	Min Poverty Index	Max Poverty Index
1	06	Epsilon High	0	1	97.2	0.8	28.3	1	28.3	28.3
2	11	Psi High	0	1	94.4	7.3	29.8	1	29.8	29.8
3	17	Lambda High	0	1	89.3	5.1	78.0	1	78.0	78.0
4	18	Delta High	0	3	94.1	2.8	82.0	3	73.3	83.0

Iteration 7 (&index=7): After the data are processed for Pi High, data set StatsForOneSchool has a total of five records.

Data Set StatsForOneSchool after Pi High Iteration: PROC PRINT Results

Obs	groupid	groupschool	_TYPE_	_FREQ_	Median Attendance Rate	Median Retention Rate	Median Poverty Index	NPoverty Index	Min Poverty Index	Max Poverty Index
1	06	Epsilon High	0	1	97.2	0.8	28.3	1	28.3	28.3
2	11	Psi High	0	1	94.4	7.3	29.8	1	29.8	29.8
3	17	Lambda High	0	1	89.3	5.1	78.0	1	78.0	78.0
4	18	Delta High	0	3	94.1	2.8	82.0	3	73.3	83.0
5	23	Pi High	0	2	91.1	6.2	80.5	2	78.0	83.0

Iteration 8 (&index=8): After the data are processed for Chi High and after eight iterations of the %DO loop, data set StatsForOneSchool has a grand total of six records, one for each high school. The resulting data set has the similar school statistics for each of the six high schools.

Data Set StatsForOneSchool after Chi High Iteration: PROC PRINT Results

Obs	groupid	groupschool	_TYPE_	_FREQ_	Median Attendance Rate	Median Retention Rate	Median Poverty Index	NPoverty Index	Min Poverty Index	Max Poverty Index
1	06	Epsilon High	0	1	97.2	0.8	28.3	1	28.3	28.3
2	11	Psi High	0	1	94.4	7.3	29.8	1	29.8	29.8
3	17	Lambda High	0	1	89.3	5.1	78.0	1	78.0	78.0
4	18	Delta High	0	3	94.1	2.8	82.0	3	73.3	83.0
5	23	Pi High	0	2	91.1	6.2	80.5	2	78.0	83.0
6	24	Chi High	0	2	92.1	3.4	80.0	2	78.0	82.0

The GroupID and GroupSchool show the school in which the corresponding group of similar schools was based on. The NpovertyIndex value shows the number of schools similar to the school specified by GroupID and GroupSchool. MinPovertyIndex and MaxPovertyIndex are the minimum and maximum poverty index values respectively in the group of similar schools to the particular school.

The example provided is simple. There are potentially more complicated situations that might arise and warrant the use of more macro parameters to facilitate processing. For example, what if the variables in the PROC MEANS VAR statement are not the same depending on the school type? What if the variables do not require medians and require means instead? Many more things can be done to embellish or make the macro more powerful.

Just imagine if one were to execute this programming situation without the use of macros. That would be a very large number of statements indeed. Macros have saved the day again in this situation. What is an otherwise daunting programming task has been reduced to a very simple one.

REFERENCES

FAQ #523: Is there a way to determine the number of observations in a data set and have that value placed in a macro variable? Retrieved November 3, 2006, from the SAS Web site: <http://support.sas.com/faq/005/FAQ00523.html>

SAS Institute Inc. 2002. *SAS[®] 9 Language Reference: Dictionary, Volumes 1 and 2*. Cary, NC: SAS Institute Inc.

TRADEMARK NOTICE

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.