

ARE YOU LOSING RECORDS & DON'T KNOW IT? TEXTMERG: A SAS® MACRO THAT HANDLES HIDDEN PROBLEMS IN MANY-TO-MANY MERGES

Debra H. Weiner, Family Health International, Research Triangle Park NC, USA

ABSTRACT

When merging data sets, one can lose records without knowing it. For example, to assign classification codes to text fields we merge a SAS data set containing a text field with a SAS data set containing classification codes for that field, merging on the text variable. We found that if (a) a single field had more than one code or (b) the same text existed in multiple records (e.g. the same health problem was reported more than once) the SAS System will not always return the correct number of records. The SAS System notes repeats of BY value but this may not alert you to a potentially serious problem.

This poster describes a macro that performs accurate merges on character BY variables. The macro can be easily modified to handle merges on numeric variables.

INTRODUCTION: THE PROBLEM

Table 1 shows selected records from a SAS data set (CRF) containing case report forms where participants in a contraceptive clinical trial were asked to state their method of birth control used immediately prior to enrollment. HIMETHOT is an open "other/specify" text field that can contain conceptually similar responses described in dissimilar ways (e.g. "The Pill", "oral contraceptive"). To categorize responses for analysis, a second data set contains each unique text string for HIMETHOT and a corresponding numeric code (CODE) and descriptor variable (DECODE). (See Table 2.) To link classification codes to individual responses you merely merge the data sets on the common text variable (HIMETHOT).

Problems arise when there are more than one record in CRF with the same value of HIMETHOT and more than one record in CODES is assigned to that value of HIMETHOT. The results of a simple merge

are given in Table 3. Note that only one CODES record is merged with each CRF record and that which record is merged is not consistent across identical cases (compare results for Participant 208 and 302). Furthermore, the problem was not consistent over the data set; for some records with the same value of HIMETHOT given in this example the merge performed correctly (data not shown). This problem also occurred when a participant had multiple records with identical text in the other/specify field (for example, multiple cases in which the same health problem was reported). The SAS System notes that the "MERGE statement has more than one data set with repeats of BY values" but this may not be enough to alert the unwary that records have been lost.

MACRO %TEXTMERG

%TEXTMERG merges two data sets on a user-specified character merge variable, outputting to a data set named in the macro call. It allows missing values on the BY variable in data set #1 (e.g. a blank text field that does not require coding). In data set #2, the BY variable must not be missing. Further, data set #2 must contain another variable, specified in the macro call (DS2VAR), which with the BY variable uniquely defines a record in that data set (e.g. CODE). The macro assumes DS2VAR is numeric but this can easily be changed. There is no maximum (other than memory capacity) on the number of multiple records in data set #2 assigned to the BY variable.

Table 4 shows the result of merging the CRF and CODES data sets using TEXTMERG.

The macro call is as follows:

```
%textmerg
(DS1, DS2, MERGVAR, DS2VAR, DS_OUT,
SORT);
```

where:

DS1 = name of first data set

DS2 = name of second data set

MERGVAR = name of character BY variable. May be missing in data set #1 but may not be missing in data set #2.

DS2VAR = name of numeric variable in the second data set that uniquely identifies a record. Must be nonmissing for all records.

DS_OUT = name you wish to assign to merged data set. It can be the same as DS1 or DS2 (though note caveats below)

SORT = sort order desired for DS_OUT. If more than one variable, separate by spaces (e.g. CENTER ID). DS_OUT will be sorted by these variables and by DS2VAR (e.g. CENTER ID CODE)

```
%macro textmerg
(DS1,DS2,MERGVAR,DS2VAR,DS_OUT,
SORT);
```

```
data cumultv;
run;
```

```
data DS1 xxx;
set &DS1;
if &mergvar = ' ' then
output xxx;
else output DS1;
run;
```

```
proc sort data=DS1;
by &mergvar; run;
```

```
data &ds2(drop=max);
set &ds2 end=last;
by &mergvar;
retain count max;
if first.&mergvar then
count=1; else count+1;
if _N_=1 then max=0;
if last.&mergvar then
max=max(count,max);
if last ne 1 then return;
else if last=1 then
call symput
('maxcount',put(max, 2.0));
run;
```

```
%do i=1 %to &maxcount;
data temp&i;
set &ds2;
if count=&i;
```

```
run;
```

```
proc sort ;
by &mergvar; run;

data crf&i;
merge DS1(in=want)
temp&i;
by &mergvar;
if want & &ds2var>.z;
run;
```

```
data cumultv;
set crf&i
cumultv;
if &ds2var>.z;
run;
```

```
%if &i=&maxcount %then %do;
data &ds_out;
set cumultv(drop=count);
run;
```

```
proc sort ;
by &sort &ds2var; run;
```

```
%end;/**if i=maxcount loop **/
%end;/**do i=1 to maxcount **/
```

```
/** merge back records with
missing text fields **/
```

```
proc sort data=xxx;
by &sort; run;
data &ds_out;
merge &ds_out xxx;
by &sort;
run;
```

```
%mend textmerg;
run;
```

COMMENTS & CAVEATS:

- The macro is designed for a character BY variable; it can easily be modified for a numeric BY variable.
- The macro is designed for a numeric DS2VAR; it can easily be modified for a character DS2VAR.
- Variables that uniquely identified a record in data set #1 no longer uniquely identify a record in the output data set. This must be taken into account in subsequent data steps.

- If the macro will be used more than once with the same data sets, separate output files are required for correct results (that is, do not assign the output data set the same name as either data set #1 or data set #2).

CONCLUSION

%TEXTMERG was created to solve a unique problem in coding text fields but either as is or with minor modifications it may be useful in other applications.

ACKNOWLEDGMENT

This paper was written under cooperative agreement CCP-A-00-95-00022-02 between Family Health International (FHI) and the US Agency for International Development (USAID). All opinions expressed are those of the author and do not necessarily reflect the views of FHI or USAID. The author thanks Angie Wheelless and Brian Calingaert for helpful comments and to the Contraceptive Research and Development Program, Arlington VA, for permission to use these data. The author, FHI, and USAID are not responsible for errors in this macro.

SAS is a registered trademark of SAS Institute Inc. in the USA and other countries.
® indicates USA registration.

Electronic copies of the macro may be obtained by sending a request via email to the author.

Debra H. Weiner
Division of Biostatistics
Family Health International
Box 13950
Research Triangle Park NC 27709
(919) 544-7040 x 348
Dweiner@fhi.org

Table 1. Selected Records from the CRF Data Set (Before Merging)

ID	HIMETHOT
101	CONDOMS & SPERMICIDE
208	MALE CONDOM & SPERMICIDE
302	MALE CONDOM & SPERMICIDE
701	CONDOMS & SPERMICIDE
1706	CONDOMS & SPERMICIDE

Table 2. Example of Multiple Coding for a Single Value of HIMETHOT, from the CODES Data Set (Before Merging)

HIMETHOT	CODE	DECODE
CONDOMS & SPERMICIDE	7	CONDOM
CONDOMS & SPERMICIDE	2	SPERMICIDE (FOAM, GEL, FILM, CREAM, SUPPOSITORIES)
MALE CONDOM & SPERMICIDE	7	CONDOM
MALE CONDOM & SPERMICIDE	2	SPERMICIDE (FOAM, GEL, FILM, CREAM, SUPPOSITORIES)

Table 3. Merged Data Set (Without Macro)

ID	HIMETHOT	CODE	DECODE
101	CONDOMS & SPERMICIDE	7	CONDOM
208	MALE CONDOM & SPERMICIDE	7	CONDOM
302	MALE CONDOM & SPERMICIDE	2	SPERMICIDE (FOAM, GEL, FILM, CREAM, SUPPOSITORIES)
701	CONDOMS & SPERMICIDE	2	SPERMICIDE (FOAM, GEL, FILM, CREAM, SUPPOSITORIES)
1706	CONDOMS & SPERMICIDE	2	SPERMICIDE (FOAM, GEL, FILM, CREAM, SUPPOSITORIES)

Table 4. Merged Data Set (Using %TEXTMERG)

ID	HIMETHOT	CODE	DECODE
101	CONDOMS & SPERMICIDE	2	SPERMICIDE (FOAM, GEL, FILM, CREAM, SUPPOSITORIES)
101	CONDOMS & SPERMICIDE	7	CONDOM
208	MALE CONDOM & SPERMICIDE	2	SPERMICIDE (FOAM, GEL, FILM, CREAM, SUPPOSITORIES)
208	MALE CONDOM & SPERMICIDE	7	CONDOM
302	MALE CONDOM & SPERMICIDE	2	SPERMICIDE (FOAM, GEL, FILM, CREAM, SUPPOSITORIES)
302	MALE CONDOM & SPERMICIDE	7	CONDOM
701	CONDOMS & SPERMICIDE	2	SPERMICIDE (FOAM, GEL, FILM, CREAM, SUPPOSITORIES)
701	CONDOMS & SPERMICIDE	7	CONDOM
1706	CONDOMS & SPERMICIDE	2	SPERMICIDE (FOAM, GEL, FILM, CREAM, SUPPOSITORIES)
1706	CONDOMS & SPERMICIDE	7	CONDOM