Paper 218-25

Crude Risk Assessment of Multi-level Exposures in Epidemiological Studies

Stuart Long, Westat, Durham, NC
Rebecca Darden, Westat, Durham, NC

Abstract

Oftentimes in epidemiological research, we deal with case-control or simple cohort studies in which the exposures have more than two levels (i.e., we are looking at an exposure by outcome table that is *rx2*). However, the CMH option in PROC FREQ will compute the estimates of risk for only 2x2 tables, and not for the more general *rx2* tables. Data sets are created for each comparison against the referent exposure group, and PROC FREQs are run on these individual data sets. The PROC FREQ then is re-run for the entire *rx2* table to obtain the estimate of the overall linear association. This can become quite a tedious undertaking very quickly. In this paper, we present macro code that automates the creation of the separate data sets and the subsequent PROC FREQs, and summarizes the cell counts and statistics of interest in an easily interpretable table for the client. The macro is written to allow use with any data set containing dichotomous outcome variables and multi-level exposure variables within any operating system. Although the macro code is advanced, use is not limited to advanced users, but, rather, to those who have additional statistical reporting needs.

Introduction

One of the initial steps in the analysis of a large case-control or simple cohort study is to look at the crude association between the exposure(s) of interest and a dichotomous outcome. The exposures may be demographic, environmental, or medical variables, and may have more than two levels. For example, a researcher may wish to look at the crude association between education level and cancer mortality, with education coded as a three-level categorical variable (<high school, high school, > high school). The Cochran-Mantel-Haenszel (CMH) statistics are often used at this point. We wish to obtain from PROC FREQ both an estimate of the overall linear association between the exposure and the outcome, and estimates of the risk at each level of the exposure.

If we have *r* levels of the exposure variable, we will have *r-1* comparisons to make against the referent group. Since the CMH option in PROC FREQ provides the risk estimates only in the case of a 2x2 table, we necessarily must create separate data sets for the *r-1* comparisons and run a PROC FREQ on each data set. Clients seldom wish to wade through the resulting stack of printouts, and a summary table is needed for them. This is a tedious exercise at best, especially when looking at several exposures each with more than two levels. Consequently, an alternative means of generating the summary table was desired. Originally the authors considered only the extraction of risk estimates for case-control studies. The macro code has been enhanced to allow the choice of case-control or cohort studies.

Coding Methods

The SAS macro code, given in Appendix 1, deals with the general case of an *rx2* table, where the value *r* is equal to the total number of levels in the exposure variable, and the outcome variable is binary. The code for this task first outputs summary statistics of the entire START data set. These statistics are loaded into macro variables, which are subsequently printed out with the final summary table. The referent level of the exposure variable, as defined by the user, is passed as a parameter to the macro and is then converted to an arbitrary value of 100 for all observations in the data set containing the referent value for the exposure variable. This value should be greater than the highest level of the exposure variable, and can be reset, by the user, to any value greater than the highest level if 100 does not suffice. The number of levels, *r*, of the exposure variable is then determined and loaded into a macro variable. Before continuing, the macro must determine which set of CMH statistics (case-control, cohort column one, cohort column two) the user has requested from the macro call. The statistical variable names are then loaded into macro variables, which will be requested when the *r-1* 2x2 tables are generated using PROC FREQ. Iterations are run at this point in the code to isolate observations containing only the referent level and one of the comparison levels of the exposure variable. During each iteration, these isolated observations are crossed with the binary outcome variable producing 2x2 tables using PROC FREQ. Each of the *r-1* iterations produces a data set with one observation that contains the requested CMH statistics for that comparison. This results in *r-1* data sets, each of which must then be identified with the comparison level of the exposure variable that was used to generate the statistics. The *r-1* data sets are then concatenated with each other to form one data set with *r-1* observations. At this point, the macro removes all generated 'one observation' data sets to prevent possible conflict with future calls of the macro in the same program or SAS session. The data set containing r-1 observations is then merged with the original START data set. Statistics will then exist for all observations in the START where the exposure variable is not equal to the referent level. Default statistics values are set for observations which contain the referent level. This data set is now ready to generate the final table using PROC TABULATE. The final table contains frequency counts with 2x2 statistics and 95% confidence intervals with p-values, as well as summary statistics from the overall linear association obtained from macro variables generated earlier in the code.

Example

In our dummy data set, we are looking at the association between Body Mass Index (BMI: weight in kilograms divided by height in meters squared) and amenorrhea. BMI is grouped into 4 levels: Lean (<19), Normal (19 to 27), Obese (28-34), and Morbid Obese (>34). Those with a normal BMI are used as the referent level.

Our outcome is a binary variable: Amenorrhea within the last year versus no amenorrhea within the last year.

With a four-level exposure variable, we would need to create *r-1* (= 3) separate data sets, run three PROC FREQs to determine the risk estimates of amenorrhea and each level of BMI, run an additional PROC FREQ to determine the overall linear association between BMI and amenorrhea, and then create a summary table.

By providing initial macro variable settings at the program's onset, the execution of this task becomes routine (see Appendix 2). Specifically, we provide to the macro the name and location of the source data set, the names of the exposure and outcome variables, the value of the exposure variable which will be the referent level, the set of statistics which the investigator desires to see in the final summary table, and the accompanying formats for those variables.

Upon execution of the code, a summary table is generated. The macro code for this example is given in Appendix 1.

Summary

The accompanying macro code alleviates the tedious programming of separate data sets and PROC FREQs when we have a case-control or simple cohort study with multi-level exposure variables. It helps provide the client with a table of pertinent summary information without overwhelming them with pages and pages of output. This macro is usable on any platform with any data set. A copy of the code in Appendix 1 and Appendix 2 can be downloaded at the following site:

http://westat.niehs.nih.gov/sugi25

Appendix 1. Macro Code

```
/* PROGRAM:  cmhmacro.sas.  This program file      */
/*                          contains all the MACRO */
/*                          code for generic       */
/*                          processing of the risk */
/*                          assessment table.      */

%MACRO cmhmacro(__dset,__exp,__outcm,__level,__stats);

/* Redirect massive data printouts to bogus file   */
/* to be erased at later time.                     */
PROC PRINTTO PRINT=garbage NEW;

/* Set the referent level to be equal to 100(or to */
/* any level that is higher than any other level   */
/* in the exposure variable)                       */
DATA cmhinput;
  SET &__dset;
  IF &__exp=&__level THEN &__exp=100;

PROC SORT DATA=cmhinput;
  BY &__exp;

/* OUTPUT statistics for the overall table of the  */
/* exposure variable with the outcome variable.    */
PROC FREQ DATA=&dset;
  TABLES &__exp*&__outcm / NOPRINT CMH;
  OUTPUT OUT=linear_k CMH NMISS N;

/* Load the p-value for table statistics into a    */
/* macro variable called __notea.                  */
/* This statistic was output from the previous     */
/* PROC FREQ.                                      */
DATA _NULL_;
  SET linear_k;
  Total_n=n+nmiss;
  CALL SYMPUT('__notea',LEFT(PUT(P_CMHCOR,4.3)));
  CALL SYMPUT('__noteb',LEFT(PUT(NMISS,5.0)));
  CALL SYMPUT('__notec',LEFT(PUT(total_n,5.0)));

/* Create an output data set that contains one     */
/* observation for each comparison level of the    */
/* exposure variable. Omit levels that = missing   */
/* (.) or = 100 (the value of the referent level). */
PROC FREQ DATA=cmhinput;
  WHERE(&__exp^=. & &__exp^=100);
  TABLES &__exp / NOPRINT OUT=levels;

/* Create the macro variable which contains the    */
/* total number of observations, n,  in the        */
/* 'levels' data set.  This will be used to        */
/* generate n-1 data sets, one for each comparison */
/* level of the exposure variable.  (A data set is */
/* not generated for the referent level, which is  */
/* why their are only n-1 data sets).              */
DATA _NULL_;
  IF 0 THEN SET levels NOBS=obs_tot;
  IF _N_=1 THEN DO;
     CALL SYMPUT('__obs_kt' , LEFT(PUT(obs_tot ,4.)));
     STOP;
  END;
RUN;

/* Create the macro variables for the statistics   */
/* selected in the MACRO call.                     */
/* Odds Ratio */
%IF &__stats=1 %THEN %DO;
    %LET __stata=_MHOR_;
    %LET __statb=L_MHOR;
    %LET __statc=U_MHOR;
    %LET __title=Mantel Haenszel Odds Ratio;
%END; %ELSE
/* Column 1 Relative Risk */
%IF &__stats=2 %THEN %DO;
    %LET __stata=_MHRRC1_;
    %LET __statb=L_MHRRC1;
    %LET __statc=U_MHRRC1;
    %LET __title=Mantel Haenszel Relative Risk;
%END; %ELSE
/* Column 2 Relative Risk */
%IF &__stats=3 %THEN %DO;
    %LET __stata=_MHRRC2_;
    %LET __statb=L_MHRRC2;
    %LET __statc=U_MHRRC2;
    %LET __title=Mantel Haenszel Relative Risk;
%END;

%MACRO runfreqs;
/* Create macro variables where each variable      */
/* contains the value for each of the comparison   */
/* levels in the exposure variable.                */
/* Iterate for each of the comparison levels in    */
/* the exposure variable.                          */
  %DO i= 1 %TO &__obs_kt;
    %GLOBAL __comp&i.;
    DATA _NULL_;
```

```
      SET levels;
      IF &i=_N_ THEN
/* Output a macro variable for the given          */
/* comparison level of the exposure variable.     */
        CALL SYMPUT("__comp&i.",LEFT(PUT(&__exp,4.)));
     RUN;

/* Run the PROC FREQ Cochran-Mantel-Haenszel (CMH) */
/* statistics for each of the exposure levels     */
/* against the referent level for the exposure    */
/* variable crossed with the outcome variable.    */
      PROC FREQ DATA=cmhinput;
        WHERE(&__exp=100 | &__exp=&&__comp&i.);
        TABLES &__exp*&__outcm / NOPRINT CMH;
        OUTPUT OUT=out&i. (KEEP = &__stata &__statb
                                  &__statc P_CMHCOR)
               CMH;
      RUN;

/* Add the value for the current comparison level  */
/* of the exposure variable back into the data set */
/* containing the CMH output statistics.          */
      DATA out&i;
        SET out&i;
        &__exp=&&__comp&i;
  %END;
  RUN;
%MEND runfreqs;

%runfreqs

/* Generate the list of valid data sets to be     */
/* included in the SET statement.                 */
%MACRO datasets;
  %DO i = 1 %TO &__obs_kt;
      out&i
  %END;
%MEND datasets;

/* Concatenate the valid data sets for each level  */
/* of the exposure variable.                      */
/* Keep the exposure variable and all pertinent    */
/* statistics for the final table.                */
DATA comps;
  SET %datasets;
RUN;

PROC SORT;
  BY &__exp;

/* Merge the CMH statistics back into the original */
/* data set. Reset the referent level from 100     */
/* back to the original value.  Set default values */
/* for the referent level statistics to be        */
/* displayed in the table.                        */
DATA allobs;
  MERGE cmhinput comps;
    BY &__exp;
  IF &__exp=100 THEN DO;
     &__exp=&__level;
     &__stata.=1;
     &__statb.=.;
     &__statc.=.;
     P_CMHCOR=.;
  END;
RUN;

/* Remove all data sets created in the macros.    */
/* This is important in order to prevent these    */
/* data sets from reappearing unexpectedly in the */
/* next exposure/outcome comparison to be run     */
/* through the macro in batch processing.         */
%DO i = 1 %TO &__obs_kt;
  PROC DATASETS memtype=data;
    DELETE out&i;
%END;
RUN;

/* Redirect the output back to the OUTPUT screen,  */
/* instead of to the file 'erase.it'.  This allows */
/* only the table to appear in the OUTPUT screen,  */
/* while all other default output from procedures  */
/* have been sent to a bogus location.            */
PROC PRINTTO;
RUN;

/* Print the table. */
PROC TABULATE DATA=allobs;
  CLASS &__exp &__outcm;
  VAR   &__stata &__statb &__statc P_CMHCOR;
  LABEL &__stata.=&__title.=
        &__statb.='Lower 95% Confidence Interval'
        &__statc.='Upper 95% Confidence Interval'
        P_CMHCOR=
           'p-value for Mantel Haenszel statistic';
  KEYLABEL MEAN=' ';
  TABLE &__exp,(all='Total N' &__outcm &__stata.*MEAN
&__statb.*MEAN
                    &__statc.*MEAN P_CMHCOR*MEAN);
  TITLE1 "OUTCM    = &__outcm";
  TITLE2 "EXPOSURE   = &__exp";
  TITLE3 "SAMPLE 'N' = &__notec. ";
  FOOTNOTE1
  "p-value for overall linear association = &__notea.";
  FOOTNOTE2
  "Compairson has &__noteb. missing values";
RUN;
%MEND cmhmacro;
```

Appendix 2. Example SAS program

```
/*************************************************/
/*  PROGRAM:   cmhtable.sas                      */
/*  AUTHORS:   Stu Long (long3@niehs.nih.gov)    */
/*             Rebecca Darden (burr@niehs.nih.gov) */
/*  DATE:      09/21/99                          */
/*  FUNCTION:  This program uses MACRO coding to  */
/*             generate an odds ratio table for a  */
/*             multi-level exposure variable with  */
/*             a dichotomous outcome variable.     */
/*************************************************/

OPTIONS NOFMTERR NOCENTER PS=48 LS=160;
LIBNAME  ahs      'd:\studies\agri\sd2';
FILENAME cmhmacro 'd:\macros\cmhmacro.sas';
FILENAME garbage  'c:\sas\erase.it';

%LET fileone =ahs.mda00301;
%LET filetwo =ahs.mda00302;
%LET filetres=ahs.mda00303;

PROC FORMAT LIBRARY=library;

%INCLUDE cmhmacro;

/*  Stats parameter: 1 = CMH odds-ratio statistics */
/*                   2 = Column 1 CMH relative     */
/*                       risk statistics           */
```

```
/*                  3 = Column 2 CMH relative     */
/*                      risk statistics           */

      /* data   exposure   outcome  referent  stats*/
      /*  set   variable   variable   level        */
      /*   ---------------------------------------*/
&cmhmacro(&fileone  ,bmi     ,amenor     ,2      ,2 );

RUN;
```

References

SAS Institute Inc. (1990). SAS/STAT User's Guide, Version 6, Fourth Edition, Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1990). SAS/MACRO LANGUAGE, Version 6, Fourth Edition, Cary, NC: SAS Institute Inc.

Acknowledgments

The authors would like to thank Marsha Shepherd, Joe Meskey, and David Shore for their assistance and review of the methods, coding, and manuscript.

Contact Information

Stuart Long, Rebecca Darden
Westat, Inc.
1009 Slater Road, Suite 120
Durham, NC 27703