Paper 282

Crude Risk Assessment of Multi-level Exposures in Case-Control Studies

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

Abstract

Oftentimes in epidemiological research, we deal with case-control 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. Datasets are created for each comparison against the referent exposure group, and PROC FREQs are run on these individual datasets. 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 datasets and the subsequent PROC FREQs, and summarizes the cell counts and statistics of interest in an easily interpretable table for the client. It is aimed at a moderate skill level; it uses macro code, PROC FREQ, and PROC TABULATE; and, though currently it is written to execute in PC SAS^{TM}, the code is not specific to an operating system.

Introduction

One of the initial steps in the analysis of a large case-control study is to look at the crude association between the exposure(s) of interest and a binary 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 datasets for the *r-1* comparisons and run a PROC FREQ on each dataset. 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.

Coding Methods

The SAS macro code, given in the Appendix 2, 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 a dataset from a PROC FREQ on the exposure variable, using a START dataset. This results in a dataset with *r* observations. This data set is used to create a macro variable containing the value of *r*. Subsequent macro variables COMPS2 to COMPS*r* are generated from this data set, where each variable contains the value of the consecutive levels of the exposure variable starting with level 2 to the final level, *r.* By invoking the WHERE statement in the PROC FREQs, we are able to generate the CMH statistics for *r-1* 2x2 tables of the exposure by outcome association. The WHERE statement allows us to keep only the observations containing either the referent level or the appropriate unique comparison level. Resulting output datasets, OUT2 to OUT*r*, contain the _MHOR_, L_MHOR, and U_MHOR statistics (the estimated Mantel-Haenszel odds ratio and associated 95% confidence interval). The exposure variable containing the comparison level used to generate each of the datasets (OUT2 to OUT*r*) is then added back to the one observation of each of the datasets for subsequent merging purposes. The datasets OUT2 to OUT*r* are concatenated to create the dataset DATASETS. The dataset DATASETS is then merged with the START dataset, by exposure. The overall linear association is output to a separate dataset that is used to load the p-value of the relationship into the macro variable NOTESTAT. PROC TABULATE uses the final merged dataset to generate a table containing the cell counts and M-H statistics for each comparison level with a footnote denoting the NOTESTAT value.

Example

In our dummy dataset, we are looking at the association between all-cancer morbidity and cigarette smoking. Smoking status can be collapsed to an ever/never variable. However, we have data on ever smoked, current smoking status, and current amount smoked; we would like to use all the available data. We categorize smoking behavior into a five-level exposure variable: Never-smoked, ex-smoker, current light smoker (<1/2 pack/day), current moderate smoker (1/2 to 1 pack/day), and current heavy smoker (>1 pack/day). Non-smokers are our referent group. Our outcome is a binary variable: Case of any cancer (except non-melanoma skin cancer), and control.

With a five-level exposure variable, we would need to create *r-1* (or 4) separate datasets, run four PROC FREQs to determine the risk estimates of cancer morbidity and each level of smoking, run an additional PROC FREQ to determine the overall linear association between smoking and cancer, 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 1). Specifically, we provide to the macro the name and location of the source dataset, the names of the exposure and outcome variables, and the accompanying formats for those variables. The exposure must be coded 0, 1, ..., r-1; the outcome must be coded with the lower number category for controls and the higher number category for cases (e.g., 1=control, 2=case). Currently, our code further requires that zero denote the referent group for the exposure, and higher levels are ordered consecutively and linearly.

Upon execution of the code, a summary table is generated. The macro code for this example is given in Appendix 2.

Summary

The accompanying macro code alleviates the tedious programming of separate datasets and PROC FREQs when we have a case-control 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. In the future, we would like to expand our code to include stratification analyses, and to streamline the existing code for easier portability.

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

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

Appendix 1. Example SAS program

```
/*******************************************************
***************************************/
/*  PROGRAM:   SUGI99.SAS
*/
/*  AUTHORS:   Stu Long / Rebecca Darden
*/
/*  DATE:      21/Sep/98
*/
/*  FUNCTION:  This program uses macro coding to /*
        generate an odds ratio table for a multi-    /*
level exposure variable with an
/*      outcome variable. */
/*******************************************************
***************************************/

OPTIONS NOFMTERR NOCENTER PS=67 LS=114
/* MLOGIC MPRINT */  ;
LIBNAME sugi     'f:\users\long\sugi99';
LIBNAME libname  'f:\users\long\sugi99';

%LET dset    =sugi.sugi;   /* name dataset */
%LET exp     =cur_smok;    /* name exposure
                              variable */
%LET outcome =event_k;     /* name outcome variable */

PROC FORMAT;
   VALUE _smoke  0 = ‘Non-smoker’
                 1 = ‘Ex-smoker’
                 2 = ‘Light smoker’
                 3 = ‘Moderate smoker’
                 4 = ‘Heavy smoker’;
   VALUE _event  1 = ‘No’
                 2 = ‘Yes’;
RUN;

%INCLUDE 'f:\users\long\sugi99\mac_lib\cmh_r2.mac';

/*******************************************************
***************************************/
```

Appendix 2. Macro Code

```
PROC SORT DATA=&dset OUT=sugi;
  BY &exp;
RUN;

/* OUTPUT statistics for all levels of exposure
variable with outcome variable              */
PROC FREQ DATA=sugi;
  TABLES &exp*&outcome / NOPRINT CMH;
  OUTPUT OUT=linear_k CMH;
RUN;

/* load p-value for linear model into a macro variable
called NOTESTAT                      */
DATA _NULL_;
  SET linear_k;
  CALL SYMPUT('notestat',LEFT(PUT(P_CMHCOR,4.3)));
RUN;

/* create an output data set that contains one
observation for each level of the            */
/* exposure variable
*/
PROC FREQ DATA=sugi;
  TABLES &exp / NOPRINT OUT=levels;
RUN;

/* create the macro variable which contains the total
number of observations in the dataset */
DATA _NULL_;
  IF 0 THEN SET levels NOBS=obs_tot;
  IF _N_=1 THEN DO;
     CALL SYMPUT('obs_kt' , LEFT(PUT(obs_tot ,8.)));
     STOP;
  END;
RUN;

/* create macro variables where each variable contains
the value for each of the comparison */
/* levels in the exposure variable
*/
%MACRO makevars;
  %DO i= 2 %TO &obs_kt;
    %GLOBAL comps&i.;
/* make the macro variables GLOBAL */
    DATA _NULL_;
      SET levels;
/* Iterate for each comparison     */
      IF &i=_N_ THEN
/* level of the exposure var.      */
         CALL SYMPUT("comps&i.",LEFT(PUT(&exp,8.)));
/* Output a macro var containing   */

/* the value of the exposure level.*/
    RUN;
  %END;
%MEND makevars;
```

```
/* Run the PROC FREQ Cochran-Mantel-Haenszel (CMH)
statistics for each of the comparison    */
/* levels against the reference level for the exposure
variable crossed with the levels of  */
/* the outcome variable
*/
%MACRO runfreqs;
  %DO I=2 %TO &obs_kt;
    PROC FREQ DATA=sugi;
      WHERE(&exp=0 | &exp=&&comps&i.);
      TABLES &exp*&outcome / NOPRINT CMH;
      OUTPUT OUT=out&i (KEEP = _MHOR_ L_MHOR U_MHOR)
CMH;
    RUN;
    DATA out&i;                               /* Add
the value for the compairson level of  */
      SET out&i;                              /*
exposure variable back into the data set   */
      &exp=&&comps&i;                         /*
containing the CMH output statistics       */
    RUN;
  %END;
%MEND runfreqs;

/* Generate the list of valid datasets for the SET
statement                                */
%MACRO datasets;
  %DO i = 2 %TO &obs_kt;
      out&i
  %END;
%MEND datasets;

%runfreqs

/* Concatenate the valid datasets for each comparison
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                                 */
DATA allobs;
  MERGE sugi comps;
    BY &exp;

RUN;

/* Print the table */
PROC TABULATE DATA=allobs;
  CLASS &exp &outcome;
  VAR   _MHOR_ L_MHOR U_MHOR ;
  LABEL _MHOR_='Mantel-Haenszel Odds Ratio'
        L_MHOR='Lower 95% Confidence Interval'
        U_MHOR='Upper 95% Confidence Interval';
  KEYLABEL mean=' ';
  TABLE &exp, (all='Total N' &outcome _MHOR_*mean
L_MHOR*mean U_MHOR*mean);
  FOOTNOTE1 "p-value for overall linear association =
&notestat ";


RUN;
```