**Program 5.5** Macro to Pool Small Missingness Pattern

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* Input parameters:

\* indata = input data set;

\* outdata = output data set;

\* varlist = a list of variables to be included in the propensity score

estimation;

\* M\_MP\_MIN = minimum number of observations for each missing pattern.

\* Missing patterns with less than MIN\_MP observations will be

pooled;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

**%MACRO** MP\_ASSIGN(MSDATA = , OUTDATA =, VARLIST =, N\_MP\_MIN = **100**);

/\* Determine how many variables to include in the propensity score

estimation \*/

%LET N = 1;

%LET VARINT = ;

%DO %UNTIL(%QSCAN(&VARLIST., &N. , %STR( )) EQ %STR( ));

%LET VAR = %QSCAN(&VARLIST. , &N. , %STR( ));

%LET VARINT = &VARINT &VAR.\*MP;

%LET N = %EVAL(&N. + 1);

%END;

%LET KO = %EVAL(&N-1);

%LET M\_MISSING = %EVAL(&N-1);

%PUT &VARINT;

%PUT &KO;

%PUT &M\_MISSING;

/\* Create indicators for missing values and missingness patterns \*/

DATA MS;

SET &MSDATA;

ARRAY MS{&M\_MISSING} M1-M&M\_MISSING.;

ARRAY X{&M\_MISSING} &VARLIST;

MV = **0**;

DO I = **1** TO &M\_MISSING;

IF X{I} = **.** THEN MS{I} = **1**;

ELSE MS{I} = **0**;

MV = **2**\*MV + MS{I};

END;

MV = MV + **1**;

DROP I;

RUN;

/\* Only keep one record for each missingness pattern \*/

PROC SORT DATA = MS OUT = PATTERN NODUPKEY;

BY MV;

RUN;

/\* Calculate the number of observations in each missingness pattern \*/

PROC FREQ DATA = MS NOPRINT;

TABLES MV / OUT = M\_MP(KEEP = MV COUNT);

RUN;

DATA PATTERN;

MERGE PATTERN M\_MP;

BY MV;

RUN;

PROC SORT DATA = PATTERN;

BY DESCENDING COUNT;

RUN;

/\* Assign missingness pattern to new index from the largest to the smallest \*/

DATA PATTERN;

RETAIN M1-M&M\_MISSING MV COUNT MV\_S;

SET PATTERN;

KEEP M1-M&M\_MISSING MV COUNT MV\_S;

MV\_S = \_N\_;

RUN;

PROC IML;

USE PATTERN;

READ ALL INTO A;

CLOSE PATTERN;

MS = A[, **1**:&M\_MISSING];

MV = A[, **1**+&M\_MISSING];

N\_MP = A[, **2**+&M\_MISSING];

MV\_S = A[, **3**+&M\_MISSING];

M\_MP = NROW(MS);

M = NCOL(MS);

/\* Calculate the distance between missingness patterns \*/

DISTANCE = J(M\_MP, M\_MP, **0**);

DO I = **1** TO M\_MP;

DO J = **1** TO I-**1**;

D = **0**;

DO L = **1** TO M;

D = D + ( (MS[I,L]-MS[J,L])\*(MS[I,L]-MS[J,L]) );

END;

DISTANCE[I,J] = D;

DISTANCE[J,I] = D;

END;

END;

I = **0**;

K\_MV\_POOL = **0**;

MV\_POOL = J(M\_MP, **1**, **0**);

/\*Pooling small missingness patterns according to their similarities to

reach a prespecified minimum number of observations (&N\_MP\_MIN) in each

pattern \*/

DO WHILE( I < M\_MP);

I = I + **1**;

IF MV\_POOL[I] = **0** THEN

DO;

K\_MV\_POOL = K\_MV\_POOL + **1**;

N\_MP\_POOL = N\_MP[I];

IF N\_MP\_POOL >= &N\_MP\_MIN THEN

DO;

MV\_POOL[I] = K\_MV\_POOL;

END;

ELSE

DO;

IF I < M\_MP THEN

DO;

A = DISTANCE[(I+**1**):M\_MP, I];

B = MV[(I+**1**):M\_MP];

C = N\_MP[(I+**1**):M\_MP];

D = MV\_S[(I+**1**):M\_MP];

E = MV\_POOL[(I+**1**):M\_MP];

TT = A || B || C || D || E;

CALL SORT( TT, {**1 3**});

J = **0**;

DO WHILE( (N\_MP\_POOL < &N\_MP\_MIN) & (I+J < M\_MP) );

J = J+**1**;

IF (TT[J,**5**] = **0**) THEN

DO;

N\_MP\_POOL = N\_MP\_POOL + TT[J,**3**];

TT[J,**5**] = K\_MV\_POOL;

END;

END;

END;

IF ( N\_MP\_POOL >= &N\_MP\_MIN ) THEN

DO;

MV\_POOL[I] = K\_MV\_POOL;

DO K = **1** TO J;

MV\_POOL[TT[K,**4**]] = K\_MV\_POOL;

END;

END;

ELSE

DO J = I TO M\_MP;

SGN\_TMP = **0**;

K = **1**;

DO WHILE(SGN\_TMP = **0** & K <= M\_MP);

DO L = **1** TO M\_MP;

IF (DISTANCE[J,L] = K) & (MV\_POOL[J]=**0**) &

(MV\_POOL[L]>**0**) THEN

DO;

MV\_POOL[J] = MV\_POOL[L];

SGN\_TMP = **1**;

END;

END;

K = K + **1**;

END;

END;

END;

END;

END;

MV\_FINAL = MV || MV\_POOL;

VARNAMES={'MV' 'MV\_POOL'};

CREATE MVPOOL FROM MV\_FINAL[COLNAME=VARNAMES];

APPEND FROM MV\_FINAL;

QUIT;

PROC SORT DATA = MVPOOL;

BY MV;

RUN;

PROC SORT DATA = MS;

BY MV;

RUN;

/\* The variable MVPOOL in the &OUTDATA set indicates the pooled missingness

pattern \*/

DATA &OUTDATA(RENAME=(MV=MP\_ORIG MV\_POOL=MP));

MERGE MS MVPOOL;

BY MV;

RUN;

**%MEND** MP\_ASSIGN;