Paper 225-25

Using SAS® Software to Perform a Case-Control Match on Propensity Score in an Observational Study

Lori S. Parsons, Cardiovascular Outcomes Research Center, Seattle, WA

ABSTRACT

In large observational studies there are often significant differences between characteristics of a treatment group and a no treatment group. Such differences should not exist in a randomized trial. These differences must be adjusted for in order to reduce treatment selection bias and determine treatment effect. There are several methods to reduce the bias of these differences and make the two groups more similar. One method is to perform analyses after matching cases (members of the treatment group) to controls (members of the no treatment group) based on a number of individual characteristics. A refinement of this method is to create a propensity score to represent the relationship between multiple characteristics and an outcome as a single score, and then match on that single score. This paper will show SAS users how to create a propensity score using the LOGISTIC procedure and then match cases to controls based on this score with a user-written SAS macro program. The results of using the presented code, run on a large observational database of myocardial infarction patients, will be given as an example.

INTRODUCTION

SAS/STAT® allows users to perform multivariate logistic regression with the LOGISTIC procedure. As with linear regression models, predicted probabilities can be obtained from logistic regression models. PROC LOGISTIC options allow users to calculate and save the predicted probability of the dependent variable, the propensity score, for each observation in the data set.

The formula for the predicted probability, $\hat{p}$, of the dependent variable is as follows:

$$\hat{p} = \frac{e^{(\hat{\alpha}+\hat{\beta}'x)}}{1 + e^{(\hat{\alpha}+\hat{\beta}'x)}}$$

where

$\hat{\alpha}$ is the intercept parameter estimate

$\hat{\beta}$ is the vector of slope parameter estimates

x is the vector of explanatory variables

This single score (between 0 and 1) then represents the relationship between multiple characteristics and the dependent variable as a single characteristic. In the case of an observational study, the dependent variable might be a treatment group. One score would be calculated for each patient in the study. This paper will show that in an observational study, cases (members of the treatment group) can be matched to controls (members of the no treatment group) on this score alone with the resulting matched populations being similar.

The data presented here are from a large observational database of myocardial infarction patients. The treatment group (N = 2,402) received an early intervention. The no treatment group (N = 17,735) did not. Table 1 includes all patients and all characteristics that were determined to relate to receiving the treatment. Differences between groups were evaluated using the rank-sum test for continuous data and the chi-squared test for binary data. For every characteristic, there was a significant difference between the treatment and no treatment groups ($p < 0.05$). For this reason, treatment selection bias needed to be reduced before outcomes could be compared.

One method to reduce this bias could have been to match cases to controls based on individual characteristics. A refinement to this method is to create a propensity score, the predicted probability of receiving the treatment, then match cases to controls based on this score. In the example presented here, a case-control match on the propensity score was performed. The results of the match are included in Table 2; seventy-eight percent of the original cases matched to a control. For the matched analysis, differences between matched pairs were evaluated using the signed rank test for continuous data and the McNemar's test for binary data.

This paper will give SAS users PROC LOGISTIC code to create the propensity score and a user-written SAS macro program to perform the case-control match. The matching macro program allows the user to choose the number of digits of the propensity score to match. A 3-digit match was performed in the example presented here.

IDENTIFY CHARACTERISTICS

The first step in creating a propensity score is to identify the characteristics, the independent variables, that relate significantly to the dependent variable. In this example, the treatment group early intervention is the dependent variable. The characteristics in Table 1 were chosen as the independent variables that relate to early intervention. It is important not to include the outcome variables of the final comparison as independent variables in the propensity score model.

CREATE THE PROPENSITY SCORE

To create the propensity score, include the selected independent variables in a PROC LOGISTIC analysis. Use the OUTPUT statement with the OUT= and PREDICTED= (or PRED= | PROB= | P=) options to create a data set that contains the predicted probability of the dependent variable for each observation. The output data set will contain all the variables from the original input data set, in addition to two new variables _LEVEL_ and the PREDICTED= variable. For binary logistic regression, the value of _LEVEL_ is equal to the value of the dependent variable being modeled and the predicted probability variable represents the probability that each observation is in the response level indicated by the _LEVEL_ variable.

The following code was run to perform the multivariate logistic regression and save the propensity score to the data set STUDY.Propen for all patients in this observational study.

```
LIBNAME STUDY 'D:\INTERVEN';
PROC LOGISTIC DATA=STUDY.SEarly Descend;
MODEL interven = ptage male white
       mhprevmi mhangina mhchf mhptca
       mhcabg mhdiab mhsmoke killip1
       pulsecd2 bpsyscd2 admitmi cpcd
       tincd
  /SELECTION = STEPWISE RISKLIMITS
       LACKFIT RSQUARE;
OUTPUT OUT= STUDY.Propen prob=prob ;
RUN;
```

OVERVIEW OF MACRO PROGRAM MATCH

In the macro program MATCH, cases will be matched to controls using the nearest available pair matching method; the cases are ordered and sequentially matched to the nearest unmatched control. If more than one unmatched control matches to a case, the control is selected at random. Randomness is achieved by generating a random number with the RANUNI function and performing a sort on that number.

A DATA step with a point command within a DO loop is used to control processing of the controls data set. This is done both for efficiency and to ensure that the same control will not be selected more than once. The controls data set is searched starting at the first unmatched control. The search stops when either a match is found, or when the propensity score in the controls data set is larger than the propensity score of the current case. If a match is found, the current search stops and the matched observation is output. The pointer is then moved forward one observation to the next unmatched observation and the next search begins from that point. If no match is found, the pointer is put back to the starting point of the last DO loop. Efficiency is achieved because the top of the controls data set is not searched again and the bottom of the data set is not searched when a potential match no longer exists. This algorithm ensures that the same control will not be selected again because the pointer is always moved past the already matched controls. Note that the point command cannot be used on a compressed data set.

The output of the MATCH macro program is a data set of matched pairs of cases and controls. One observation will be output for each matched case and control. The output data set will also contain the variable *matchto*. *Matchto* is used to identify pairs. It is a numeric field with potential values 1 to n, where n is the total number of cases. Each pair will have the same value.

MACRO PROGRAM MATCH

```
%MACRO MATCH (
  Lib,        /* Library Name          */
  Dataset,    /* Data set of all       */
              /* patients              */
  Matched,    /* Data set of Matched   */
              /* Pairs                 */
  SCase,      /* Sorted data set of    */
              /* Cases                 */
  SControl,   /* Sorted data set of    */
              /* Controls              */
  depend,     /* Dependent variable    */
              /* that indicates        */
              /* Case or Control;      */
              /* Code 0 for Cases,     */
              /* 1 for Controls.       */
  digits      /* Format of the number  */
              /* of digits to match    */
  );
/*
```

Part 1: Create Case and Control Data Sets

The first part of the macro program creates separate data sets for cases and controls and performs appropriate sorts. A random number is generated in the controls data set. The case data set is sorted in propensity score order. The control data set is sorted in propensity score then random number order. Including the random number ensures that, if there is more than one control that could be matched to a case, the match will be a random selection from all identical controls. The indicator variable *Cmatch* will later keep track of matches; it is initialized to 0=No. The variables *aprob* and *cprob* are the propensity score rounded to the selected number of digits.

```
*/
data tcases (drop=cprob)
     tctrl  (drop=aprob) ;
   set &LIB..&dataset. ;
   /* Create the data set of Controls*/
   if &depend. = 1 and prob ne .
   then do;
      cprob  = round(prob,&digits.);
      Cmatch = 0;
      Length RandNum 8;
      RandNum=ranuni(1234567);
      Label RandNum=
         'Uniform Randomization Score';
      output tctrl;
    end;
   /* Create the data set of Cases   */
   else if &depend. = 0 and prob ne .
   then do;
      aprob =round(prob,&digits.);
      output tcases;
   end;
run;
```

```
proc sort data=tcases
   out=&LIB..&Scase.;
   by prob;
run;
proc sort data=tctrl
   out=&LIB..&Scontrol.;
   by prob randnum;
run;

/*
```

Part 2: Perform the Match

The next part of the macro program performs the match and outputs the matched pairs. First, the cases data set is selected. *Curob* is used to keep track of the current case. *Matchto* is used to identify matched pairs of cases and controls. *Start* and *oldi* are initialized to control processing of the controls data set DO loop.

```
*/
data &lib..&matched.
   (drop=Cmatch randnum aprob cprob start
         oldi curctrl matched);
   set &lib..&SCase. ;
   curob + 1;
   matchto = curob;

   if curob = 1 then do;
      start = 1;
      oldi = 1;
   end;
/*
```

Next, the controls data set is selected. Processing starts at the first unmatched observation. The data set is searched until a match is found, or it is determined no match can be made. Error checking is performed to avoid an infinite loop. *Curctrl* is used to keep track of current control.

```
*/
   DO i = start to n;
      set &lib..&Scontrol. point = i nobs = n;

      if i gt n then goto startovr;
      if _Error_ = 1 then abort;

      curctrl = i;
/*
```

If the propensity score of the current case (*aprob)* matches the propensity score of the current control (*cprob),* then a match was found. Update *Cmatch* to 1=Yes. Output the control. Update *matched* to keep track of last matched control. Exit the DO loop. If the propensity score of the current control is greater than the propensity score of the current case, then no match will be found for the current case. Stop the DO loop processing.

```
*/
      if aprob = cprob then
      do;
         Cmatch = 1;
         output &lib..&matched.;
         matched = curctrl;
         goto found;
      end;
      else if cprob gt aprob then
         goto nextcase;

    startovr: if i gt n then
       goto nextcase;
   END; /* end of DO LOOP  */
```

```
/*
```

If no match was found *(Cmatch=0)*, do not increment the controls starting place of the next DO loop; put the pointer back to the starting place of the last DO loop (*start = oldi).* If a match was found (*Cmatch=1*), increment the controls starting place of the next DO loop by 1 (*oldi = matched + 1* and *start = matched + 1*). Output the matched case. In both events, retain the DO loop processing fields.

```
*/
   nextcase:
   if Cmatch=0 then start = oldi;

   found:
   if Cmatch = 1 then do;
      oldi = matched + 1;
      start = matched + 1;
     set &lib..&SCase. point = curob;
        output &lib..&matched.;
   end;

   retain oldi start;
   if _Error_=1 then _Error_=0;
run;
%MEND MATCH;
```

MACRO MATCH CALL STATEMENT

The following are call statements to the macro program MATCH. The first performs a 4-digit match; the second performs a 3-digit match.

```
%MATCH(STUDY,Propen,Match4,SCase4,
       SContrl4,Interven,.0001);

%MATCH(STUDY,Propen,Match3,SCase3,
       SContrl3,Interven,.001);
```

RESULTS

Table 2 shows the results after matching cases to controls based on a 3-digit match of the propensity score. Differences between matched pairs were evaluated using the signed rank test for continuous data and the McNemar's test for binary data. For every characteristic, there is no longer a significant difference between the treatment and no treatment groups ($p < 0.05$).

CONCLUSION

By using a combination of the SAS/STAT LOGISTIC procedure and a user-written SAS macro program, cases can be matched to controls on propensity score alone in observational studies. Such a match will result in two groups with similar characteristics, thus reducing selection bias. Outcomes can then be compared between the two similar groups. This paper has given SAS users code to create the propensity score and then match the cases to the controls based on this single score. Real data has been used to illustrate the differences in the original study population and similarities in the resulting matched population. A limitation to this method is

how well your data are fit in the multivariate model from which the propensity score was computed. And, as with matching on individual characteristics, this method can only reduce bias in measured characteristics.

CONTACT INFORMATION

Contact the author at:

Lori S. Parsons
Cardiovascular Outcomes Research Center
1910 Fairview Ave. E. Suite 205
Seattle, WA 98102
Work Phone: (206) 720-4453
Fax: (206) 521-1690
Email: lparsons@u.washington.edu

REFERENCES

Connors AF Jr., Sperroff T, Dawson N, Thomas C, Harrell FE Jr., Wagner D, Desbiens N, Goldman L, Wu AW, Califf RM, Fulkerson, WJ Jr., Vidaillet H, Broste S, Bellamy P, Lynn J, Knaus WA for the SUPPORT Investigators (1996), "The Effectiveness of Right Heart Catheterization in the Initial Care of Critically Ill Patients," *Journal of the American Medical Association*, 276:889-897.

Rosenbaum PR. (1989), "Optimal Matching for Observational Studies," *Journal of the American Statistical Association*, 84:1024-1032.

Rubin DB. (1997), "Estimating Causal Effects from Large Data Sets Using Propensity Scores," *Annals of Internal Medicine*, 127:757-763.

SAS Institute Inc. (1995), *Logistic Regression Examples Using the SAS® System, Version 6, First Edition,* Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1989), *SAS/STAT® User's Guide, Version 6, Fourth Edition*, Volume 1, Cary, NC: SAS Institute Inc.

Table 1: Initial Comparison of All Patients

	Early Intervention N	(%)	Conservative N	(%)	p-value
Total Patients	2,402		17,735		
Age					
Mean ±s.d.	61.3 ±12.2		68.2 ±13.0		<0.0001
25th Percentile	52.1		59.0		
50th Percentile	61.6		69.4		
75th Percentile	71.7		78.1		
N Missing	4		14		
Gender					
Female	658	(27.4%)	6,821	(38.5%)	
Male	1,744	(72.6%)	10,914	(61.5%)	<0.0001
Race					
Non-White	185	(8.2%)	1,959	(11.6%)	
White	2,079	(91.8%)	15,002	(88.4%)	<0.0001
Medical History:					
Angina					
No	1,958	(81.5%)	13,294	(75.0%)	
Yes	444	(18.5%)	4,441	(25.0%)	<0.0001
Previous MI					
No	1,828	(76.1%)	12,353	(69.7%)	
Yes	574	(23.9%)	5,382	(30.3%)	<0.0001
CHF					
No	2,302	(95.8%)	15,174	(85.6%)	
Yes	100	(4.2%)	2,561	(14.4%)	<0.0001
CABG					
No	2,024	(84.3%)	14,423	(81.3%)	
Yes	378	(15.7%)	3,312	(18.7%)	0.0005
PTCA					
No	2,045	(85.1%)	15,797	(89.1%)	
Yes	357	(14.9%)	1,938	(10.9%)	<0.0001
Diabetes					
No	2,008	(83.6%)	12,840	(72.4%)	
Yes	394	(16.4%)	4,895	(27.6%)	<0.0001
Smoking					
No	1,586	(66.0%)	13,097	(73.8%)	
Yes	816	(34.0%)	4,638	(26.2%)	<0.0001
Presentation:					
Heart Failure					
No CHF	2,089	(87.5%)	13,210	(75.0%)	
Rales, JVD	226	(9.5%)	2,956	(16.8%)	<0.0001
Pulmonary edema	73	(3.1%)	1,437	(8.2%)	<0.0001
Pulse > 100					
No	2,096	(88.0%)	13,075	(74.8%)	
Yes	285	(12.0%)	4,404	(25.2%)	<0.0001
Systolic BP <= 100					
No	2,159	(90.3%)	16,481	(94.3%)	
Yes	232	(9.7%)	1,003	(5.7%)	<0.0001
Presentation Chest Pain					
No	175	(7.4%)	3,097	(17.9%)	
Yes	2,180	(92.6%)	14,164	(82.1%)	<0.0001
Admission Dx MI					
No	1,330	(56.1%)	14,869	(85.1%)	
Yes	1,040	(43.9%)	2,598	(14.9%)	<0.0001
Transferred-in					
No	1,972	(82.1%)	12,657	(71.4%)	
Yes	430	(17.9%)	5,078	(28.6%)	<0.0001

Table 2: Comparison of Matched Patients

	Early Intervention N	(%)	Conservative N	(%)	p-value
Total Patients	1,882		1,882		
Age					
Mean ±s.d.	62.3 ± 12.0		62.2 ± 13.4		0.9606
25th Percentile	53.4		51.8		
50th Percentile	62.9		62.2		
75th Percentile	71.6		72.3		
N Missing	0		0		
Gender					
Female	557	(29.6%)	522	(27.7%)	
Male	1,325	(70.4%)	1,360	(72.3%)	0.2072
Race					
Non-White	159	(8.4%)	151	(8.0%)	
White	1,723	(91.6%)	1,731	(92.0%)	0.6353
Medical History:					
Angina					
No	1,514	(80.4%)	1,511	(80.3%)	
Yes	368	(19.6%)	371	(19.7%)	0.9020
Previous MI					
No	1,423	(75.6%)	1,409	(74.9%)	
Yes	459	(24.4%)	473	(25.1%)	0.5971
CHF					
No	1,790	(95.1%)	1,788	(95.0%)	
Yes	92	(4.9%)	94	(5.0%)	0.8805
CABG					
No	1,576	(83.7%)	1,562	(83.0%)	
Yes	306	(16.3%)	320	(17.0%)	0.5400
PTCA					
No	1,607	(85.4%)	1,605	(85.3%)	
Yes	275	(14.6%)	277	(14.7%)	0.9266
Diabetes					
No	1,549	(82.3%)	1,514	(80.4%)	
Yes	333	(17.7%)	368	(19.6%)	0.1429
Smoking					
No	1,268	(67.4%)	1,253	(66.6%)	
Yes	614	(32.6%)	629	(33.4%)	0.6032
Presentation:					
Heart Failure					
No CHF	1,640	(87.1%)	1,640	(87.1%)	
Rales, JVD	184	(9.8%)	172	(9.1%)	0.5459
Pulmonary edema	58	(3.1%)	70	(3.7%)	0.2982
Pulse > 100					
No	1,633	(86.8%)	1,629	(86.6%)	
Yes	249	(13.2%)	253	(13.4%)	0.8479
Systolic BP <= 100					
No	1,730	(91.9%)	1,715	(91.1%)	
Yes	152	(8.1%)	167	(8.9%)	0.3801
Presentation Chest Pain					
No	156	(8.3%)	153	(8.1%)	
Yes	1,726	(91.7%)	1,729	(91.9%)	0.8586
Admission Dx MI					
No	1,190	(63.2%)	1,202	(63.9%)	
Yes	692	(36.8%)	680	(36.1%)	0.6845
Transferred-in					
No	1,531	(81.3%)	1,527	(81.1%)	
Yes	351	(18.7%)	355	(18.9%)	0.8674