Pairwise matching of sample and control groups using PROC FASTCLUS

Kathleen A. Travis, University of Iowa, Iowa City, IA
Stephan Amdt, University of Iowa, Iowa City, IA
Nancy C. Andreasen, University of Iowa, Iowa City, IA

ABSTRACT

In many biomedical studies it is desirable to match control subjects to patient subjects on several variables (e.g. age, sex, and education) while studying the effects of some treatment. If performed by hand, this matching process can require many hours of work for a data set of moderate to large size. One automated solution to this problem is to utilize the FASTCLUS procedure within SAS®. Individuals from the group to be matched to are assigned as "seeds" and individuals from the second group are allowed to form clusters around the seeds based upon the values of the variables being matched on. This allows the match for a given seed to be obtained by choosing from the cluster a subject with variable values which minimize the sum of squared distances from the cluster seed values. This paper presents the SAS software codes to implement this procedure and the application is discussed relative to the use of large archival data sets which are often used in biomedical settings.

This program utilized the Base SAS Software and SAS/STAT® Software. The program is intended for beginning and intermediate users.

INTRODUCTION

In some studies, patient and control subjects are "matched" to obtain experimental groups which are similar with respect to specific variables such as age, IQ or gender. The goal may be to obtain matched pairs of individuals or to obtain groups which are not significantly different on some variable or combination of variables. One of the reasons for matching is to remove possible sources of error such as those which result from cohort effects (2). For example, elderly persons are likely to be shorter than younger persons based upon the nutritional levels and medical care available during the period in which each was growing. Misleading results in comparing the effect of any random variable on height would occur if one group was predominantly aged while the other was predominantly young. In addition, if paired or dependent t-tests will be used for analysis, the power of the tests will be increased as a function of the relevance of the matching variables.

The perfect match would be one in which a subject from group one had variable values which were identical to those values for a subject from group two. However, it happens rather infrequently that identical matches on a number of variables can be obtained. It becomes necessary to match subjects as closely as possible by trying to minimize the differences between them for each variable. When matching to a large archival data set, this can become an extremely time consuming task if performed by hand. One automated solution is to utilize the FASTCLUS procedure within SAS. PROC FASTCLUS relies upon a general clustering technique called nearest centroid sorting (3). Individuals from the group to be matched to are assigned as "seeds" and individuals from the second group are allowed to form clusters around the seeds by minimizing the within group (cluster) sum of squares (1,4,5).

Frequently, categorical variables (sex, handedness, diagnosis, etc.) need
to be exactly matched between groups while continuous variables can be allowed to exhibit a range of acceptable values between the matched pair or group. The SAS program presented here matches strictly on categorical variables and allows equal weighting of continuous variables. This involves a two step process. The seed and cluster groups are initially segregated on the basis of the categorical variable (e.g. sex). The first run of the program forms clusters for only one value of the categorical variable (e.g. sex=male). The FASTCLUS procedure is used to assign cluster members around the seeds. This approach is advantageous for several reasons. Since for categorical variables, only those cluster members of the correct category would be considered, the amount of data to be manipulated is reduced by this initial screening, thus increasing the efficiency of the program. In addition, having members of the wrong category included in the data run, will cause different clusters to be formed, thus, providing less optimum choices for seed matches.

APPLICATION

In the data sets below, information on individuals is stored at a central SAS library referred to as central. For each individual, sociodemographic information such as age at visit, IQ and sex are stored in a data set called master. The purpose of this demonstration program was to match a group of controls to a seed group of patients on the variables age, sex and IQ.

In the data step, input the patient "seed" file. Included in the file are an identifying variable (id) and 3 variables which will be matched on (age, sex, and IQ). For the first run, only the data for males are kept at the input step. In this example, data on control subjects was used to form clusters around the patient seeds. An array of the data from the patient and control groups is formed using age and IQ as criteria. The data for these two criteria is standardized to a mean of 0 and a variance of one.

The options for the FASTCLUS procedure are given in the proc statement. Seed=seeds and data=pool specify data sets to be used. The list option causes the program to list all observations, the values of the id the number of the cluster to which the observations is assigned, and the distance between the observation and the seed. Since we wish to specify the patient data as the seeds and not allow replacement of seeds, the replace=none option is chosen. In addition, since there will be no replacement of seeds, the maxiter (maximum iteration number) option is set to 0.

CODE

options nocenter linesize=78;
Libname here 'disk:\[home\]';
Libname core 'central:\[archive.sasdata\]';

Input your seed data file, keeping only the males for the first run.

Note that sex=1 are males and sex=2 are females

;**************
data mine;
infile here.pat;
input id $IQ age sex;
if sex = 1;
   keep id IQ age sex;
proc sort;
   by id;
proc print;
   var id IQ age sex;
   title 'Original male sample to match';
From the archive, save out the variables that you wish to match on.

For the first run keep only the male controls. Note that control=1 are controls, control=0 are non-controls.

In the proc sort data=mine statement, change the file name as needed.

```
data master;
set archive.master;
if control=1 and sex=1;
keep id IQ sex age;
proc sort;
by id;
```

In the master (in=them) mine statement, change the data set name.

```
data all;
set master (in=them) mine;
by id;
if them then case = 1;
else case = 0;
array criter age IQ;
array new nage niq;
do over criter;
    new = criter;
end;
```

```
proc standard mean =0 sId =1;
var nage niq ;
```

```
data pool seeds;
set all;
if case then do;
    output pool;
end;
else do;
    output seeds;
end;
```

```
proc fastclus seed=seeds data=pool list replace=none maxiter=0
out= specifies an output data set which will contain the original data, plus the new variables cluster and distance.
maxcluster= specify the number of seed subjects in your input statement.
```

```
data list;
set magic seeds (in = seedfile);
if seedfile then distance =0;
```

Change the var names to those that you are matching on.

for seed= specify the name of the data set containing your seeds

for data= specify the name of the data set containing the subjects who will form the cluster

list causes the program to list all observations, the values of the id, the number of the cluster to which the observations is assigned, and the distance between the observation and the seed

replace=none suppresses seed replacement.

maxiter=0 specifies that there will be no iterations for recomputing cluster seeds

```
proc contents data = work.magic;
```
proc sort;
  by cluster distance;

**********************************************************************
In the print statement below, change the var names to those that you are matching on (e.g., age IQ sex).
**********************************************************************

proc print;
  id id;
  var distance age 10 sex;
  by cluster;
  title 'Male Matches';
proc contents;

**********************************************************************
Rerun the program changing the categorical variable sex to sex=2 for female matches.
**********************************************************************

SAMPLE OUTPUT

In the output shown below the values for the variables to matched on are printed for the original male sample "seeds".

<table>
<thead>
<tr>
<th>OBS</th>
<th>ID</th>
<th>AGE</th>
<th>IQ</th>
<th>SEX</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>20</td>
<td>90</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>25</td>
<td>95</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>30</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>30</td>
<td>110</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>22</td>
<td>120</td>
<td>1</td>
</tr>
</tbody>
</table>

Cluster 1 is formed around the values for ID 1 in the seed file shown above. Note that for ID 21 the distance measure is 0, it is an exact match for ID 1. Other cluster members are listed in order of increasing distance from the seed.

Cluster=1

<table>
<thead>
<tr>
<th>ID</th>
<th>DISTANCE</th>
<th>AGE</th>
<th>IQ</th>
<th>SEX</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>0.00000</td>
<td>20</td>
<td>90</td>
<td>1</td>
</tr>
<tr>
<td>101</td>
<td>1.00000</td>
<td>20</td>
<td>91</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>5.00000</td>
<td>23</td>
<td>86</td>
<td>1</td>
</tr>
<tr>
<td>111</td>
<td>7.81025</td>
<td>25</td>
<td>84</td>
<td>1</td>
</tr>
<tr>
<td>211</td>
<td>7.81025</td>
<td>25</td>
<td>84</td>
<td>1</td>
</tr>
</tbody>
</table>

Cluster 2 formed around ID 2 in the seed file. Note that there are no exact matches for this cluster and that ID 26 is the closest match.

Cluster=2

<table>
<thead>
<tr>
<th>ID</th>
<th>DISTANCE</th>
<th>AGE</th>
<th>IQ</th>
<th>SEX</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>2.0000</td>
<td>25</td>
<td>97</td>
<td>1</td>
</tr>
<tr>
<td>143</td>
<td>2.2361</td>
<td>23</td>
<td>96</td>
<td>1</td>
</tr>
<tr>
<td>110</td>
<td>3.0000</td>
<td>25</td>
<td>92</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>3.1623</td>
<td>24</td>
<td>92</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>3.1623</td>
<td>26</td>
<td>92</td>
<td>1</td>
</tr>
<tr>
<td>172</td>
<td>3.6056</td>
<td>22</td>
<td>97</td>
<td>1</td>
</tr>
<tr>
<td>162</td>
<td>4.4721</td>
<td>21</td>
<td>97</td>
<td>1</td>
</tr>
<tr>
<td>201</td>
<td>5.6569</td>
<td>29</td>
<td>91</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>7.2111</td>
<td>19</td>
<td>99</td>
<td>1</td>
</tr>
<tr>
<td>121</td>
<td>7.6158</td>
<td>32</td>
<td>92</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>8.5440</td>
<td>33</td>
<td>92</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>8.9443</td>
<td>29</td>
<td>87</td>
<td>1</td>
</tr>
<tr>
<td>182</td>
<td>10.8167</td>
<td>19</td>
<td>104</td>
<td>1</td>
</tr>
<tr>
<td>141</td>
<td>11.4018</td>
<td>32</td>
<td>86</td>
<td>1</td>
</tr>
</tbody>
</table>

LIMITATIONS

1. The FASTCLUS procedure is intended for use with large data sets, from approximately 100 to 100,000 observations. With small data sets, the results may be highly sensitive to the order of the observations in the data set(1).
2. No seed may have a missing data point for any variable.
3. The program does not guarantee that a cluster will be formed around every seed. This is likely to occur when one or more of the seed variables
lie at the extreme of the range for that variable.

(4) On occasion, a poor match with a subject from the archive will be obtained. For example, if distances for var1-var3 are 1 each and distance for var4 is 4 this would be listed as a better match than if distances for each of the 4 variables was 2. This problem can be obviated by using a cutoff for individual variables. For example, no single variable can have a distance of 4 or more and still be considered a good match.

REFERENCES


ACKNOWLEDGEMENTS

This research was supported in part by NIMH Grants MH31593, MH40856 and MHCR43271.

Kathleen A. Travis, Ph.D.
Univ. of Iowa Hospitals and Clinics
Mental Health Clinical Research Center
2911 JPP
200 Hawkins Drive
Iowa City, Iowa, 52242
(319) 353-4942
kathleen-travis@uiowa.edu