Paper 1861-2014

CASE CONTROL MATCHING: COMPARING SIMPLE DISTANCE- AND PROPENSITY SCORE-BASED METHODS

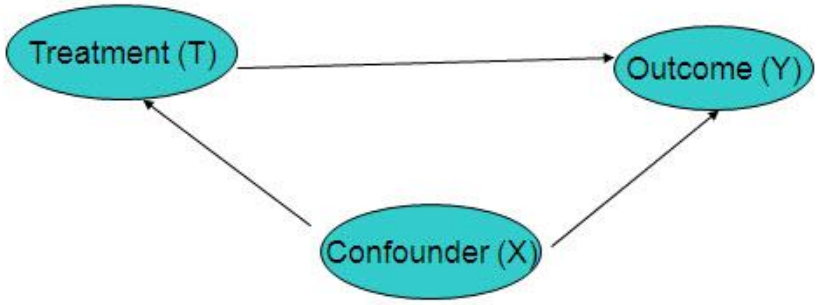
Lovedeep Gondara, BC Cancer Agency; Colleen McGahan, BC Cancer Agency

ABSTRACT

A case control study is basically comparing a case series to a matched control series and is commonly implemented in the field of public health. While matching is intended to eliminate confounding, the main potential benefit of matching in case control studies is a gain in efficiency. There are many known methods for selecting potential match or matches (in case of 1:n studies) per case, the most prominent being distance-based approach and matching on propensity scores. In this paper, we will go through both and compare their results .

INTRODUCTION

Case control study can be simply defined as a form of retrospective observational study where people with a certain condition or treatment are compared to people without that condition to study the differences.



The basic idea behind matching is to remove confounding and also to gain efficiency. In other words, the main goal of matching is to increase the study's efficiency by forcing the case and control samples to have similar distributions across confounding variables. Whereas a careful consideration is required to choose potential confounders for minimizing any bias, this is not the topic we will focus on.

DIFFERENT APPROACHES

1. Individual paired matching using multivariates

In individual paired matching for each case, one or more controls with relevant matching criteria are chosen.

It can be exact – Example: All males are matched to males or stage 3 cancer patients are matched to stage 3 cancer patients.

It can have a range - Example: For continuous variables like age and weight calipers (±5) can be used.

Methods proposed such as

Euclidean distance

Dij= SQRT [SUM { W.k(X.ik-X.jk)**2}]*

Or

Weighted sum of absolute differences

*Dij= SUM { W.k*ABS(X.ik-X.jk) }*

Where

the sum is over the number of matching factors X(with index k) and

W.k = the weight assigned to matching factor k and
X.ik = the value of variable X(k) for subject i.

Or

Mahalanobis Distance

This can be calculated between two column vectors as

$md(X_i, X_j) = \{ (X_i - X_j)'S^{-1}(X_i - X_j)\}^{1/2}$

where S is the sample covariance matrix of X.

can be used to compute the distance.

This type of matching is done on P-levels of space where P is number of variables selected to be matched on.

2. **Propensity Score Matching (PSM)**

Propensity score is defined as the conditional probability of receiving a particular treatment given other covariates. It was introduced by Rosenbaum[1] and Rubin[1] in 1983 and now is the most used technique for bias reduction of observational data. In SAS® propensity scores can be estimated using PROC LOGISTIC. Then similar distance measures can be used to find potential controls.

Various algorithms can be used to choose a match, but the most common and widely implemented is the Nearest available Match using Greedy approach. This approach suggests that once a match is found within specified caliper, it is never broken even if there may exist a better or closer match. Even though the optimal matching technique guarantees the best and closest match possible, It is more computationally intensive and does not guarantee much improvement over Greedy approach.

CASE STUDY

We have a sample of 43 cases receiving a certain type of treatment for breast cancer and a random sample of 1991 controls from general Breast cancer population. The Table below shows the comparison of both cohorts before matching. A modified version of gmatch[2] is used which is capable of matching on propensity scores and Mahalanobis distance with propensity score calipers.

Variable Name	Variable description	Pvalue
Age	Age at time of diagnosis	<.001
Days to Tx	Days to treatment from diagnosis	0.08
Size of lesion	Size in cm	0.5
ER Status	Estrogen Receptor Status	<0.001
Grade	Grade of tumor	0.05
Number of Positive Nodes	Number of nodes positive	<0.001

Table 1. Before Matching

As it is clear from above table both samples are unbalanced on prognostic factors like age, ER status and Grade.

1:1 match using individual multivariate paired matching.

Macro Call –

```
%match_mod(data=data,group=group, id=id,
           var=age days size ER Grade ,weight=1 1 1 1 1 ,caliper=5 60 1 0 0,
           type=0, ncontls=1,out=matched);
```

Where

Data= name of dataset with cases and controls

Group= group variable with 1 for case and 0 for control

Id= id variable

Var = variables used for matching

Weight= Non negative weights for each variable

Caliper= range of values to match on

Type= Type of matching (0- multivariate, 2 – Propensity score with user defined caliper, 3 – Nearest neighbor match using propensity score, 4 – Mahalanobis distance using propensity score as caliper)

Ncontls= number of controls per case

Out= Output Dataset

Variable Name	**Pvalue**
Age	<.17
Days to diagnosis	0.80
Size of lesion	0.21
ER Status	0.48
Grade	0.55
Number of Positive Nodes	0.05

Table 2. Multivariate matching results

Methodology: Weighted sum of absolute differences(SAD) is used to find matching controls. Trial and error approach can be used for obtaining closest possible match per variable by adjusting the calipers at every run and looking at the output. Often the researcher has to make choice between quality and quantity of matches as the similarity between matched pairs often decrease with an increase in the number of matches.

It can be calculated in SAS as

```
%do k=1 %to &nvar;
%scan(&wts,&k)*abs(__ca&k - __co&k )
%if &k<&nvar %then + ;
%end;
```

Where nvar is number of variables, wts is the weight supplied and ca and co are case and control respectively.

1:1 matching using propensity score with calipers

Macro Call –

```
%match_mod(data=data,group=group, id=id,
           Class=ER Grade, var=age days size, weight=1, caliper=0.1,
           type=2, ncontls=1,out=matched);
```

Variable Name	Pvalue
Age	0.52
Days to diagnosis	0.84
Size of lesion	0.13
ER Status	0.03
Grade	0.83
Number of Positive Nodes	0.46

Table 3. Propensity score with calipers matching results

Methodology: Propensity scores are calculated by running logistic regression and calipers can be adjusted to obtain suitable matches. Often a good balance can be achieved easily as one does not have to manipulate calipers for each variable but just for the propensity score.

Propensity scores can be calculated in SAS using PROC LOGISTIC as

```
proc logistic data=data;
class  ER Grade;
model group= age days size ER Grade;
output out=ps_p XBETA=ps_xb STDXBETA= ps_sdxb PREDICTED = ps_pred;
run;
```

Where ps_p is the name of output dataset which has propensity scores by name of ps_pred which can be used to calculate distance between cases and controls using the same method as above.

1:1 propensity score match(Nearest Neighbor)

```
%match_mod(data=data,group=group, id=id,
           Class=ER Grade, var=age days size, weight=1, type=3,
           ncontls=1,out=matched);
```

Variable Name	Pvalue
Age	<.008
Days to diagnosis	0.15
Size of lesion	0.46
ER Status	0.01
Grade	0.02
Number of Positive Nodes	0.66

Table 4. Propensity score with NN match results

Methodology: Nearest neighbor (NN) match always guarantees matches for all cases as there is no limit imposed upon difference in propensity scores to select a match. In this method, both treatment and control units are first randomly sorted. Then the first treatment unit is selected to find its closest control match based on the absolute value of the difference between the propensity score (or the logit of the propensity score) of the selected treatment and that of the control under consideration. The closest control unit is selected as a match. This procedure is repeated for all the treated units.

Implementation in SAS

```
ScoreDistance = abs(scoretrt - scorecs);
if ScoreDistance < BestDistance then do;
BestDistance = ScoreDistance;
Control_ID= id_1;
Case_ID= id_2;
```

Where Bestdistance can be set to a large value (e.g. 99), scoretrt and scorecs are propensity scores for treatment and control group respectively and id_1 and id_2 are id's for cases and controls respectively.

Mahalanobis metric matching using propensity score as calipers

```
%match_mod(data=data,group=group, id=id,
           Class=ER Grade, var=age days size, weight=1 1 1 1 1 1, type=4,
           ncontls=1,out=matched);
```

Variable Name	Pvalue
Age	0.56
Days to diagnosis	0.21
Size of lesion	0.98
ER Status	0.79
Grade	0.61
Number of Positive Nodes	0.78

Table 5. Mahalanobis metric matching results

Methodology: This is a more robust type of matching , in which one treatment group can be successfully matched to another treatment group using Mahalanobis metric matching within the calipers defined by the propensity score method. This has been shown in previous studies to be a superior matching technique than the rest above.

Implementation in SAS

See Wuwei Wayne Feng[3] et al. for detailed implementation.

CONCLUSION

This paper illustrated various case control matching techniques and there comparison. Although there is enough study and evidence pointing to use of Mahalanobis metric matching using propensity score caliper's as a more robust matching method and often resulting in two treatment groups having very similar baseline characteristics. But other matching techniques can be optimized for similar results as well. For true bias reduction thorough understanding and identification of true confounders is must.

REFERENCES

- Carpter, R. G. Matching when covariables are normally distributed. Biometrika, 64, 299 – 307, 1977
- [3]Wuwei Wayne Feng, Yu Jun, Rong Xu. A Method/Macro Based on Propensity Score and Mahalanobis Distance to Reduce Bias in Treatment Comparison in Observational Study
- [2]Erik Bergstralh & Jon Kosanke. Computerized matching of cases to controls using the greedy matching algorithm with a fixed number of controls per case (Macro gmatch)
- Cochran, W. G. and Rubin, D. B. "Controlling bias in observational studies: A review". Sankhya, Ser. A. 35 (1973), 417-446
- Lori S Parsons, "Using SAS software to perform a case-control match on propensity score in an observational study", SUGI 1999, 225-229,
- Lori S. Parsons, "Reducing Bias in a propensity score matched-pair sample using greedy matching techniques", Proceeding of the twenty-fifth annual SAS users group internal conference, Cary, NC: SAS Intistitue Inc. 2000, 1166-1171.

- [1]Robenbaum, P. R. and Rubin, D. B. “The central role of the propensity score in the observational studies for causal effects’, Biometrika, 70, 41—55 (1983).
- Robenbaum, P. R. and Rubin, D. B. “Constructing a control group using multivariate matched sampling methods that incorporate the propensity score”, The American Statistician, 39, 33-38 (1985) .
- Rubin, D. B. Matching to remove bias in observational studies: a review. Sankhya, series A 35, 417-446. 1973a
- Rubin, D.B. “Bias reduction using Mahalanobis metric matching” Biometrics, 36, 293-298 (1980).
- Rubin. D.B. “Estimating causal effects from large data sets using propensity score”. Annals of Internal medicine, October, 1997, 127: 757-763.
- Sekhon, Jasjeet S. 2007. “Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R “ Journal of Statistical Software.
- Marcelo Coca-Perraillon. “Matching with Propensity Scores to Reduce Bias in Observational Studies”

ACKNOWLEDGMENTS

Thanks to Anky Lai of BC Cancer Agency for her help in reviewing this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Lovedeep Gondara
BC Cancer Agency
686 west broadway
Vancouver BC
Lovedeep.gondara@bccancer.bc.ca