Application of Propensity Score – an overview with the help of r software package

1. Introduction

Experimental studies or randomized clinical trial in health care setting are usually the preferred type of research when we want to compare to treatments or two groups or test procedures. When the experimental or clinical trials are not feasible or unethical to conduct then observational studies are preferred over the experimental studies to compare the two groups out of which one is received the treatment under consideration and another one is not the received the treatment or placebo group. One of the main drawbacks of the observational studies is that the experimental units are not randomized among the two groups which results in imbalance in the base line covariates which ultimately affect the outcome under consideration which is not the case in experimental or clinical trials. Propensity score[1,2] helps us to overcome the imbalance in the two groups by assigning propensity scores to treatment groups and non-treatment groups and the members of the two groups with equal propensity scores are matched which makes the two groups comparable with respect to the covariates. This paper provides an overview of propensity score, its application and computing the propensity score with the use of R statistical package which is an open source statistical package.

Keywords: Propensity Score, matching, observational studies, r package

2. Propensity Score

Propensity score is computed as a conditional probability of experiment unit is assigned to treatment group given the covariates. The covariates from the study are selected for calculating the propensity score which have an effect on the outcome. The covariate selection can be done by the referring to the similar studies [3].

After calculating the propensity score, matching of experimental units in the two groups with respect to the propensity score can be done by the following algorithms [1, 4,5]:

a. Nearest Neighbor algorithm
b. Manhattan algorithm
c. Caliper matching algorithm
After matching the two groups, the outcome measure is computed to find out the treatment effect [6,7,8].

3. **Advantages of Propensity score matching** [9,10,11,12]  
   1. To help to balance the treatment and non-treatment group and to study treatment effect on the outcome  
   2. Reduces selection or allocation bias  
   3. Reduces bias in the outcome estimation

4. **Disadvantages of Propensity score matching** [1,2]  
   1. It considers only observed covariates while computing the propensity score

5. **Propensity Score computation example**

   Let us consider a hypothetical observational study which aims to study to observe the effect of treatment (patients who took aspirin) and non-treatment group (who have not taken aspirin) on the outcome of heart attack. Here the two groups of patients are not randomly assigned and hence the outcome heart attack might be affected by the covariates such as age, gender, food habits, diabetic status, hypertension status, smoking and alcohol status.

   The propensity score is computing by calculated by conditional probability of patient being in the aspirin group given the above covariates like age, gender etc.

   After matching the two groups of patients, the outcome measure is computed to find out the effect of Aspirin on heart attack.

   The data set used in the study can be downloaded from the site: [www.ijsmi.com/book.php](http://www.ijsmi.com/book.php)

6. **R software package for computing Propensity score**

   MatchIt R software[13] can be used to compute the Propensity score. The following code needs to be executed directly in the environment or in the R Studio[14] environment which is Integrated Development Environment for R package.
R code for computing Propensity Score

#install the required package using install.package statement
Install.packages("MatchIt")
#call the MatchIt package into R studio environment using library statement
library(MatchIt)
#call the readxl package using library statement to import excel data sheet
library(readxl)
#Data clinical1.xlsx can be downloaded from website www.ijsmi.com/book.php
clinical1 <- read_excel("C:/test/cclinical1.xlsx")
#we use the following covariates age, diet, bilirubin and creatine for the test purpose
ps = matchit(group ~ age + diet + bilirubin + creatinine, data = clinical1, method = "nearest")
#we generate the output using summary() and plot() statements
summary(ps)
plot(ps)
newclinial1 <- match.data(ps)

7. Results

The following output is obtained using the above code

Table-1 : Summary of Balance for all the data

<table>
<thead>
<tr>
<th></th>
<th>Means Treated</th>
<th>Means Control</th>
<th>SD Control</th>
<th>Mean Diff</th>
<th>eQQ Med</th>
<th>eQQ Mean</th>
<th>eQQ Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>distance</td>
<td>0.5642</td>
<td>0.4077</td>
<td>0.1681</td>
<td>0.1564</td>
<td>0.1656</td>
<td>0.1651</td>
<td>0.2516</td>
</tr>
<tr>
<td>age</td>
<td>42.9655</td>
<td>36.129</td>
<td>10.5916</td>
<td>6.8365</td>
<td>8</td>
<td>7.4483</td>
<td>12</td>
</tr>
<tr>
<td>diet</td>
<td>1.6552</td>
<td>1.5161</td>
<td>0.508</td>
<td>0.139</td>
<td>0</td>
<td>0.1379</td>
<td>1</td>
</tr>
<tr>
<td>bilirubin</td>
<td>2.8931</td>
<td>2.3871</td>
<td>1.3423</td>
<td>0.506</td>
<td>0.6</td>
<td>0.6379</td>
<td>1.5</td>
</tr>
<tr>
<td>creatinine</td>
<td>1.0138</td>
<td>0.9161</td>
<td>0.582</td>
<td>0.0977</td>
<td>0.1</td>
<td>0.1483</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Table-2: Summary of balance for matched data

<table>
<thead>
<tr>
<th></th>
<th>Means Treated</th>
<th>Means Control</th>
<th>SD Control</th>
<th>Mean Diff</th>
<th>eQQ Med</th>
<th>eQQ Mean</th>
<th>eQQ Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>distance</td>
<td>0.5642</td>
<td>0.4256</td>
<td>0.1586</td>
<td>0.1386</td>
<td>0.1432</td>
<td>0.1394</td>
<td>0.2195</td>
</tr>
<tr>
<td>age</td>
<td>42.9655</td>
<td>36.6552</td>
<td>10.6176</td>
<td>6.3103</td>
<td>8</td>
<td>6.4483</td>
<td>11</td>
</tr>
<tr>
<td>diet</td>
<td>1.6552</td>
<td>1.5517</td>
<td>0.5061</td>
<td>0.1034</td>
<td>0</td>
<td>0.1034</td>
<td>1</td>
</tr>
<tr>
<td>bilirubin</td>
<td>2.8931</td>
<td>2.4241</td>
<td>1.3179</td>
<td>0.469</td>
<td>0.5</td>
<td>0.5655</td>
<td>1.4</td>
</tr>
<tr>
<td>creatinine</td>
<td>1.0138</td>
<td>0.9552</td>
<td>0.5804</td>
<td>0.0586</td>
<td>0.1</td>
<td>0.1276</td>
<td>0.4</td>
</tr>
</tbody>
</table>
Table-3: Percent Balance Improvement

<table>
<thead>
<tr>
<th></th>
<th>Mean Diff.</th>
<th>eQQ Med</th>
<th>eQQ Mean</th>
<th>eQQ Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>distance</td>
<td>11.4298</td>
<td>13.5016</td>
<td>15.5269</td>
<td>12.773</td>
</tr>
<tr>
<td>age</td>
<td>7.6961</td>
<td>0</td>
<td>13.4259</td>
<td>8.3333</td>
</tr>
<tr>
<td>diet</td>
<td>25.6</td>
<td>0</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>bilirubin</td>
<td>7.3203</td>
<td>16.6667</td>
<td>11.3514</td>
<td>6.6667</td>
</tr>
<tr>
<td>creatinine</td>
<td>39.9772</td>
<td>0</td>
<td>13.9535</td>
<td>20</td>
</tr>
</tbody>
</table>

Sample sizes:

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Treated</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>31</td>
<td>29</td>
</tr>
<tr>
<td>Matched</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td>Unmatched</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Discarded</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Here we can see that only 2 patients are matched and no patient is discarded with respect to the covariates and the balance between the two experiment groups is introduced. The mean difference between the treatment group and non-treatment group (control group) was higher before the balancing and after balancing it was improved as shown in the Table-1, Table-2 and Table-3.

For example the mean difference between the treatment group and control group was initially 6.83 and after balancing it was reduced to 6.31 and the percentage improvement was 7.69.

The following plot is obtained from the above output.
The balanced data set can be used for further analysis using `match.data` statement.

**Conclusion**

The paper provided an overview of propensity score with the help of R software package using hypothetical example. The propensity matching introduced balance between the treatment and the control group.
References

13. https://cran.r-project.org/