# **Propensity Score Analysis to Match Intervention Kids to Non-Intervention**

## **Introduction:**

It is often unethical to use random assignments in areas where important research or constructs are being carried out. Propensity Score Analysis is a method of reducing selection bias in estimating treatment effects by reducing many pre-treatment covariates to a single scalar function and matching and allows researchers to compare subjects with similar probability to receive the treatment. Our analysis is based on the assumption that students are chosen for a particular treatment based on their background characteristics.

We use the propensity score to model the relationship between pretreatment variables and treatment assignment which is the conditional probability of assignment to treatment based on measured pretreatment analysis. In our analysis, we have used the birth age of the individual, age of the individual at removal, congregate care type, number of placements, number of removals, sex and race as our pre-treatment covariates to match the kids who have received treatment to those who have not received treatment.

## Procedure followed:

The code is written in R using the MatchIt package and follows the flowchart as shown below:

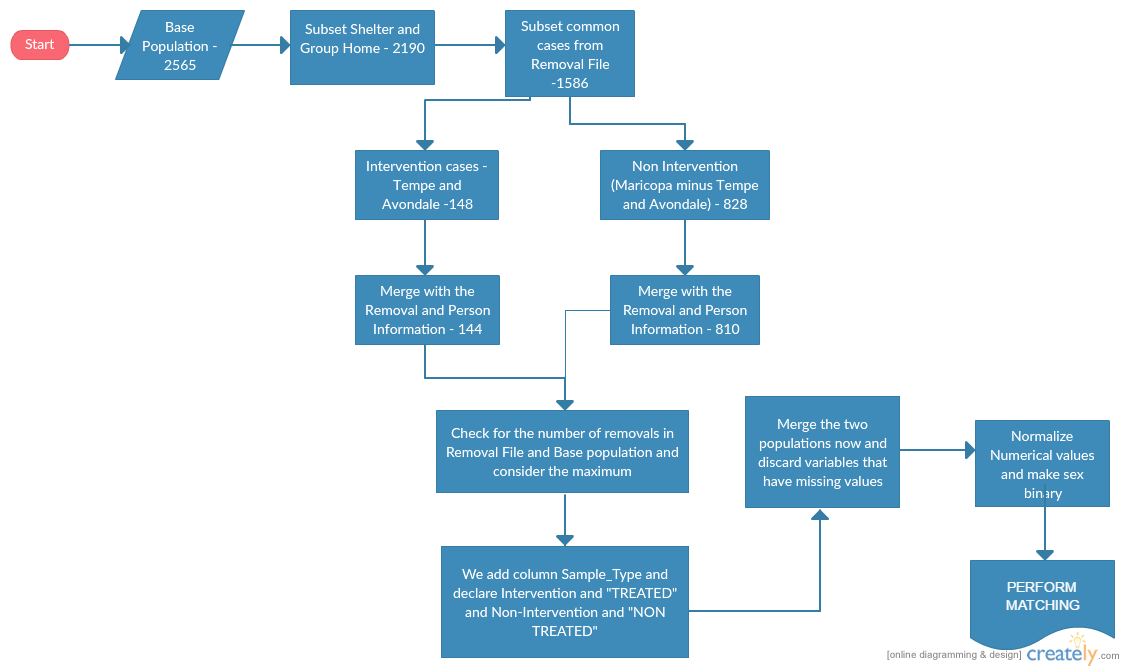


Figure 1:PSM flowchart

The various steps taken can be described in detail as given below:

1. We initially load all the raw data files in R as needed for processing. Below are the different files loaded with their names we used for convention and the names in the bracket are the actual raw file names located in [Raw Data-Path](file:///Z:\TITLE_IV_E_WAIVER_DCS\OUTCOME_EVALUATION\DATA\2016\Raw_Data):
2. The Base\_Pop file (ASU AFCARS Base-Pop.csv)
3. The Removal\_File (RPT187\_REMOVAL\_FILE\_20170106\_09231498.csv)
4. Person\_File (RPT187\_PERSON\_FILE\_20170106\_09231498.csv)
5. Placement\_File(RPT187\_PLACEMENTS\_FILE\_20170106\_09231498.csv)
6. We set the date columns given in each of these files to the proper date format so that we can access them as dates.
7. The Base Population initially has 2565 child entries, out of which we filter the “Shelter” and “Group Home” because “Institution/RTC” will not be considered. This leaves us with 2190 entries in the Base\_Pop file which we now call Base\_Pop2.
8. We then match on the Case Ids given in the resultant Base\_Pop2 File to the Case Id column in the Removal\_File so that we have the subset of kids on which we can have all the removal information. Thus after finding the common cases that number to 1586, we obtain Base\_Pop3 with 1586 cases.
9. We now divide these cases in the Base\_Pop into intervention kids and non-intervention kids from Maricopa based on the unit numbers associated with them. According to the information we have, only the kids from Tempe and Avondale underwent intervention. We make vectors of unit numbers belonging to different regions of Maricopa county and then subset for appropriate regions to obtain the intervention and non-intervention data.
   1. So we subset and store the kids from Tempe and Avondale into another file that we call Base\_Pop4 which now has 148 cases.
   2. The second subset consists of all other kids whose unit numbers are associated with Maricopa county but do not belong to Tempe and Avondale. This becomes the non-intervention Base\_Pop5 subset from Maricopa having 828 cases.
10. Both of the files – Base\_Pop4 and Base\_Pop5 now have to be merged with the Removal and Person information so as to enable us to build variables for both the files. This is done in two steps:
11. The Removal File is merged with the Person File to include all Removals and matching the child ids in in both the files. We call the merged file as Removal\_Person\_merger. We then code the PERS.GENDER attribute in this file as follows- Male (M) is coded as 1 and Female(F) is coded as 2. This is done so that we are able to match this column to the gender column in Base\_Pop files.
12. We now merge the Removal\_Person\_merger with Base\_Pop4 and Base\_Pop5 to obtain Base\_Pop4\_Removal and Base\_Pop5\_Removal based on the following attributes – Case ID, Latest Removal date, Date of Birth and the Gender. We perform an inner join so that only the matching values from both the merging parties are included in the resultants and there exist no NAs. This is because matching cannot be performed on missing values. So now the Base\_Pop4\_Removal has 144 cases and Base\_Pop5\_Removal has 810 cases.
13. We add the variable ‘number of removals’ for every child as a separate variable. While checking, we noticed that the number of removals information obtained from the Base\_Pop file is different from that obtained from the Removal File. In this case we take the maximum of these two numbers is the correct one assuming that in either of the files the extra removal information might not have been recorded.
14. We now create the two populations for matching by selecting appropriate variables that we want to match on and also labelling them as treatment and non-treatment. We introduce a new variable Sample type in each Base\_Pop4 and Base\_Pop5 and enter ‘Treated’ for the intervention population and ‘Non-Treated’ for the non-intervention population.
15. These two dataframes are now merged into one called ‘Population’ vertically and create a Boolean variable called GROUP that’ll be true for Treated and false for non-treated.
16. We discard all variables that have missing values as they will not aid in matching. This leaves us with the variables – Sex, Races,