Analysis of the NHANES data using propensity score adjustment

July 26, 2017

Abstract

The file contains the analysis of the NHANES data from R using various propensity score adjustment methods.

1. Data

```
rm(list=ls())
file.remove(list.files(pattern='.pdf'))

+  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
+ [15] TRUE TRUE TRUE

list.of.packages <- c("NHANES", "tableone", "Matching", "MatchIt","survey",
                      "twang","SuperLearner","glmnet","polspline","randomForest","SIS","Hmisc")
new.packages <- list.of.packages[!(list.of.packages %in% installed.packages()[,"Package"])]
if(length(new.packages)) install.packages(new.packages,verbose=FALSE)

library(NHANES,verbose=FALSE)
library(SuperLearner,verbose=FALSE)
library(nnet,verbose=FALSE)
library(nnls,verbose=FALSE)
library(glmnet,verbose=FALSE)
library(polspline,verbose=FALSE)
library(randomForest,verbose=FALSE)
library(SIS,verbose=FALSE)
library(twang,verbose=FALSE)
library(tableone,verbose=FALSE)
library(survey,verbose=FALSE)
library(Matching,verbose=FALSE)
library(MatchIt,verbose=FALSE)
library(Hmisc,verbose=FALSE)
```

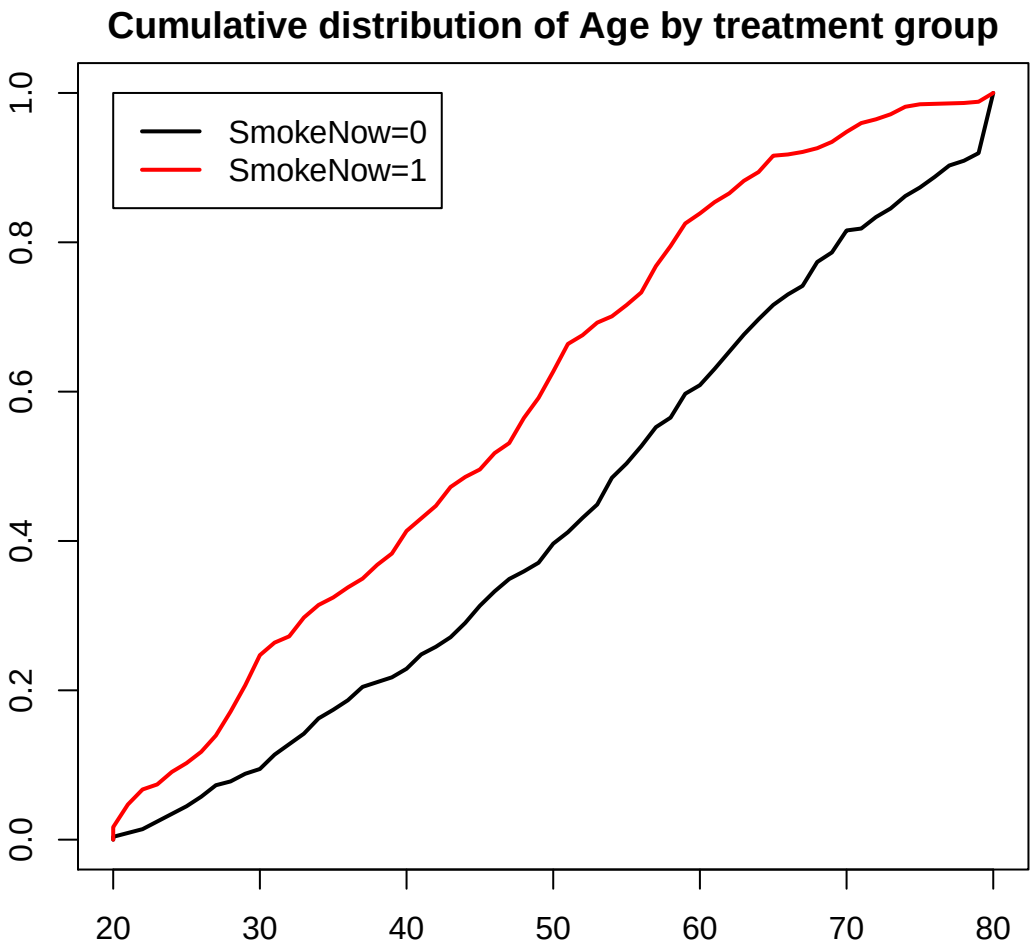
```
set.seed(37)
NHANES$SmokeNow <- as.numeric(NHANES$SmokeNow)-1
small.nhanes <- na.omit(NHANES[NHANES$SurveyYr=="2011_12" & NHANES$Age > 17,c(3,4,8:11,13,25,61)])
dim(small.nhanes)

+ [1] 1377    9

names(small.nhanes)

+ [1] "Gender"        "Age"           "Race3"         "Education"
+ [5] "MaritalStatus" "HHIncome"      "Poverty"       "BPSysAve"
+ [9] "SmokeNow"
```

```
temp.mat <- model.matrix(~Gender*Age+Gender*Race3+Gender*Education+Gender*MaritalStatus+
                  Gender*HHIncome+Gender*Poverty+Age*Race3+Age*Education+Age*MaritalStatus+
                  Age*HHIncome+Age*Poverty+Race3*Education+Race3*MaritalStatus+Race3*HHIncome+
                  Race3*Poverty+Education*MaritalStatus+Education*HHIncome+Education*Poverty+
                  MaritalStatus*HHIncome+MaritalStatus*Poverty+HHIncome*Poverty,data=small.nhanes)

interact.data <- model.matrix(~Gender*Age+Gender*Race3+Gender*Education+Gender*MaritalStatus+
                  Gender*HHIncome+Gender*Poverty+Age*Race3+Age*Education+Age*MaritalStatus+
                  Age*HHIncome+Age*Poverty+Race3*Poverty+Education*Poverty+MaritalStatus*Poverty+
                  HHIncome*Poverty,data=small.nhanes)
interact.data <- data.frame(interact.data)
interact.data$SmokeNow <- small.nhanes$SmokeNow
vars.interact <- colnames(interact.data)[30:107]
```

```
vars <- c("Gender", "Age", "Race3", "Education", "MaritalStatus", "Poverty")
tabUnmatched <- CreateTableOne(vars = vars, strata = "SmokeNow", data = small.nhanes, test = FALSE)
print(tabUnmatched, smd = TRUE)
```

```
+                      Stratified by SmokeNow
+                       0             1             SMD
+   n                     782           595
+   Gender = male (%)     432 (55.2)    369 (62.0)   0.138
+   Age (mean (sd))     54.33 (16.52) 44.96 (15.11)  0.592
+   Race3 (%)                                        0.315
+      Asian               25 ( 3.2)     15 ( 2.5)
+      Black               43 ( 5.5)     64 (10.8)
+      Hispanic            26 ( 3.3)     38 ( 6.4)
+      Mexican             45 ( 5.8)     35 ( 5.9)
+      White              630 (80.6)    416 (69.9)
+      Other               13 ( 1.7)     27 ( 4.5)
+   Education (%)                                    0.512
+      8th Grade           59 ( 7.5)     33 ( 5.5)
+      9 - 11th Grade      71 ( 9.1)    120 (20.2)
+      High School        152 (19.4)    151 (25.4)
+      Some College       256 (32.7)    210 (35.3)
+      College Grad       244 (31.2)     81 (13.6)
+   MaritalStatus (%)                                0.488
+      Divorced            85 (10.9)     77 (12.9)
+      LivePartner         61 ( 7.8)     96 (16.1)
+      Married            453 (57.9)    240 (40.3)
+      NeverMarried       108 (13.8)    142 (23.9)
+      Separated            6 ( 0.8)     14 ( 2.4)
+      Widowed             69 ( 8.8)     26 ( 4.4)
+   Poverty (mean (sd))  3.11 (1.65)   2.38 (1.58)   0.453
```

```
temp0 <- Ecdf(small.nhanes$Age[small.nhanes$SmokeNow==0],pl=F)
temp1 <- Ecdf(small.nhanes$Age[small.nhanes$SmokeNow==1],pl=F)
par(mar=c(2,3,2,1))
plot(temp0$x,temp0$y,ylab="ECDF(Age)",xlab="Age",main="",type="l",lwd=2)
lines(temp1$x,temp1$y,col="red",lwd=2)
title('Cumulative distribution of Age by treatment group')
legend(20,1,c('SmokeNow=0','SmokeNow=1'),col=c('black','red'),lty=1,lwd=2)
```

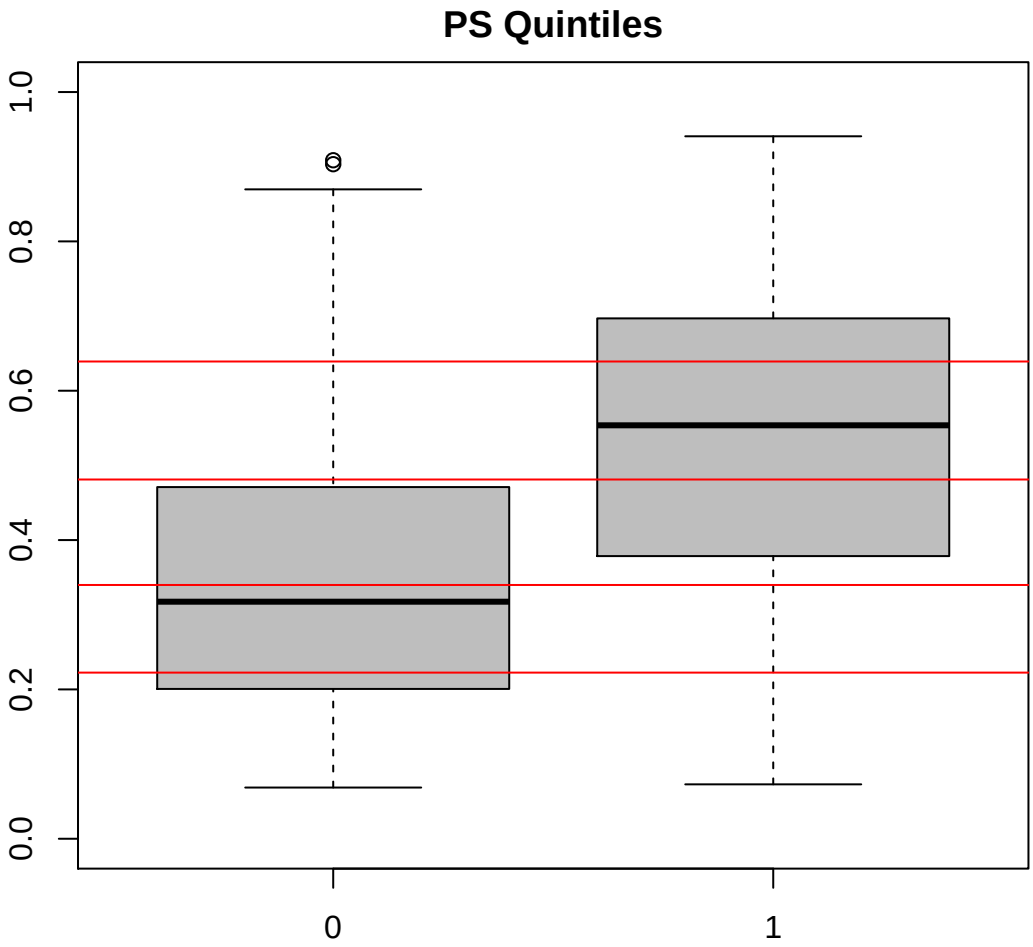



```
##########
## on interactions?
SMDinteract <- CreateTableOne(vars = vars.interact, strata = "SmokeNow", data=interact.data,test = FALSE)
summary(ExtractSmd(SMDinteract))

+     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
+ 0.003419 0.037239 0.094537 0.133918 0.172850 0.587127
```

2. Logistic regression

```
ps.mod <- glm(SmokeNow~Gender+Age+Race3+Education+MaritalStatus+HHIncome+Poverty,
              data=small.nhanes,family="binomial")
ps.lr <- predict(ps.mod,type="response")

small.nhanes$ps.lr <- ps.lr
(quints <- c(0,quantile(ps.lr,seq(.2,1,.2))))

+                 20%       40%       60%       80%      100%
+ 0.0000000 0.2224824 0.3398586 0.4811227 0.6391357 0.9407413

summary(ps.lr)

+    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
+ 0.06858 0.25228 0.40382 0.43210 0.59621 0.94074
```

```
Smoke<-small.nhanes$SmokeNow
par(mar=c(2,3,2,1))
boxplot(ps.lr[Smoke==0],ps.lr[Smoke==1],main='PS Quintiles',
             ylab="PS",xlab="Treatment Group",names=c(0,1),ylim=range(0,1),col='gray')
abline(h=quints[2:5],col="red")
```



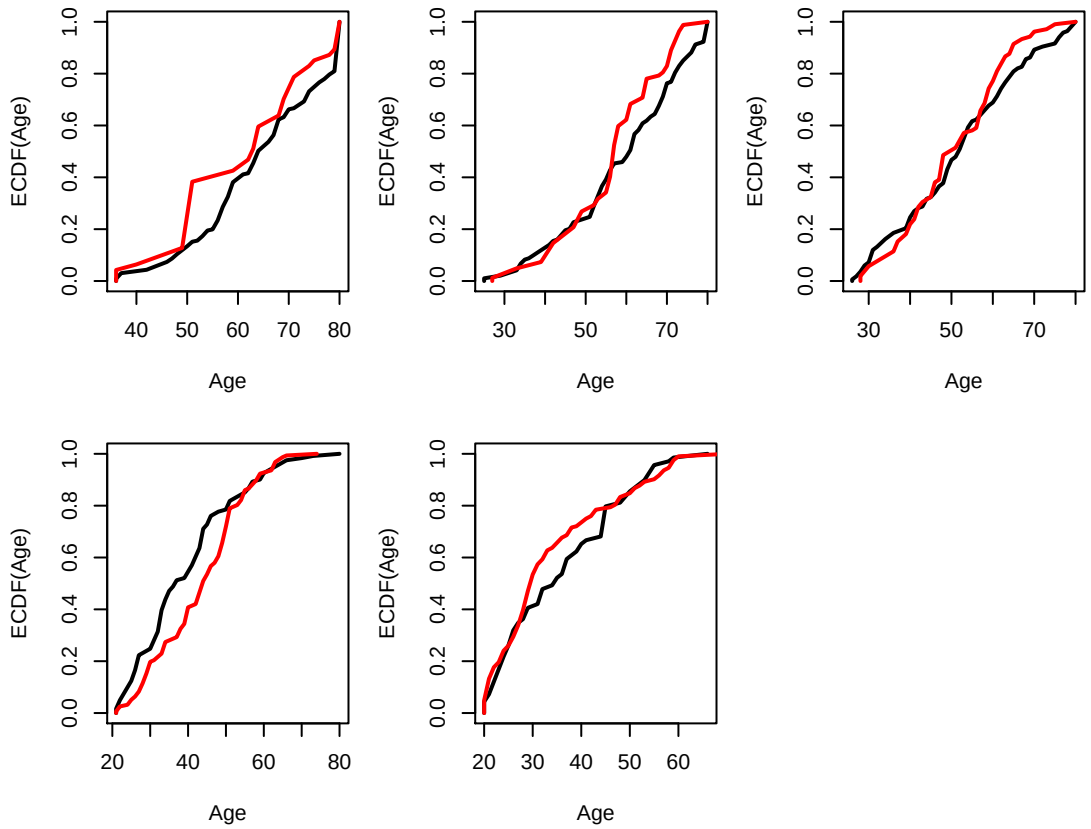

```
small.nhanes$ps.lr <- ps.lr
rbind(table(cut(ps.lr[small.nhanes$SmokeNow==0],quints)),
      table(cut(ps.lr[small.nhanes$SmokeNow==1],quints)))

+      (0,0.222] (0.222,0.34] (0.34,0.481] (0.481,0.639] (0.639,0.941]
+ [1,]       231          194          167           121            69
+ [2,]        47           82          105           157           204
```

```
ps.lr.quints <- cut(ps.lr,quints,labels=1:5)
small.nhanes$ps.lr.quints <- ps.lr.quints

par(mfrow=c(2,3),mar=c(4,4,2,1))
for(j in 1:5) {
  nonsmj <- small.nhanes$Age[small.nhanes$SmokeNow==0 & small.nhanes$ps.lr.quints==j]
  temp0 <- Ecdf(nonsmj,pl=FALSE)
  smj <- small.nhanes$Age[small.nhanes$SmokeNow==1 & small.nhanes$ps.lr.quints==j]
  temp1 <- Ecdf(smj,pl=FALSE)
  plot(temp0$x,temp0$y,ylab="ECDF(Age)",xlab="Age",main="",type="l",lwd=2)
  lines(temp1$x,temp1$y,col="red",lwd=2)
}
```

```
SMD.table <- ExtractSmd(tabUnmatched)
for(j in 1:5) {
  tabPSquints <- CreateTableOne(vars = vars, strata = "SmokeNow",
                                data = small.nhanes[ps.lr.quints==j,], test = FALSE)
  SMD.table <- cbind(SMD.table,ExtractSmd(tabPSquints))
}
round(SMD.table,3)
```

```
+               SMD.table
+ Gender            0.138 0.102 0.104 0.029 0.200 0.031
+ Age               0.592 0.257 0.171 0.099 0.311 0.164
+ Race3             0.315 0.317 0.112 0.344 0.415 0.287
+ Education         0.512 0.538 0.417 0.280 0.238 0.302
+ MaritalStatus     0.488 0.432 0.239 0.272 0.233 0.261
+ Poverty           0.453 0.087 0.126 0.114 0.004 0.146
```

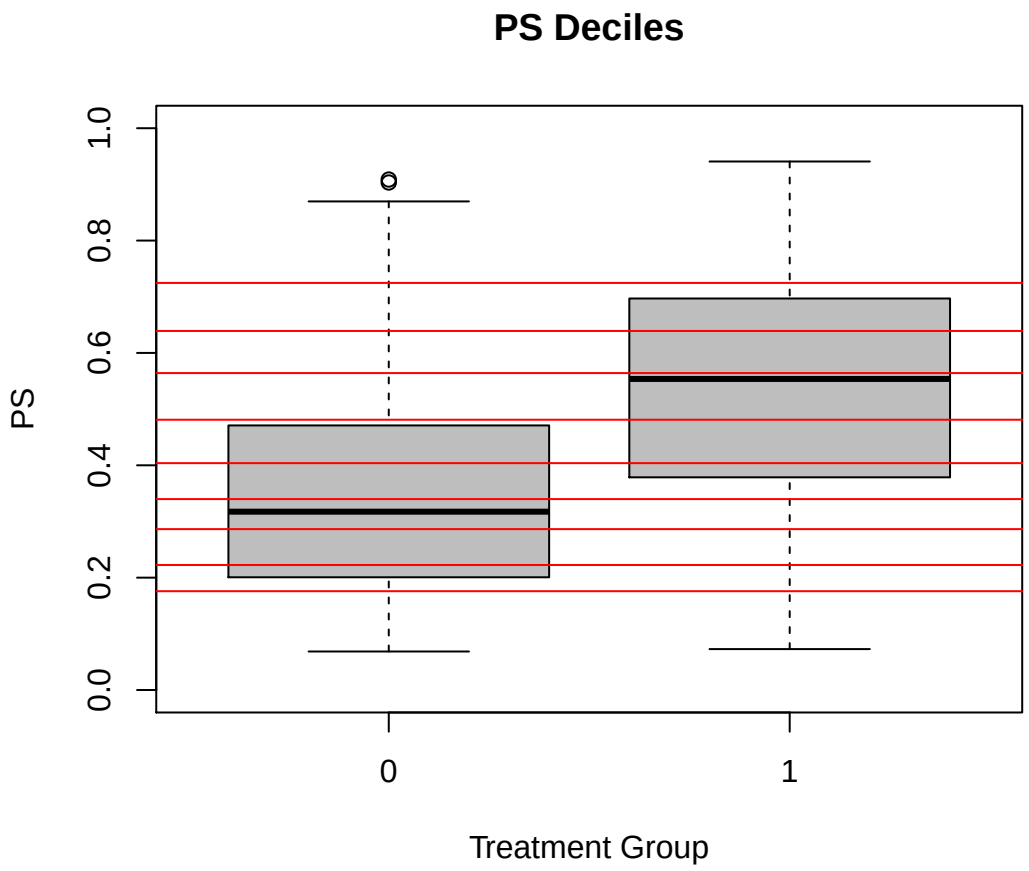
```
Max.SMD <- max(SMD.table);Mean.SMD <- mean(SMD.table); Med.SMD <- median(SMD.table)
```




```
boxplot(ps.lr[small.nhanes$SmokeNow==0],ps.lr[small.nhanes$SmokeNow==1],ylim=range(0,1),
            ylab="PS",xlab="Treatment Group",names=c(0,1),col='gray')
(dec <- c(0,quantile(ps.lr,seq(.1,1,.1))))
```

```
+                 10%       20%       30%       40%       50%       60%
+ 0.0000000 0.1758902 0.2224824 0.2863810 0.3398586 0.4038208 0.4811227
+       70%       80%       90%      100%
+ 0.5641231 0.6391357 0.7246066 0.9407413
```

```
abline(h=dec[2:10],col="red");title('PS Deciles')
```

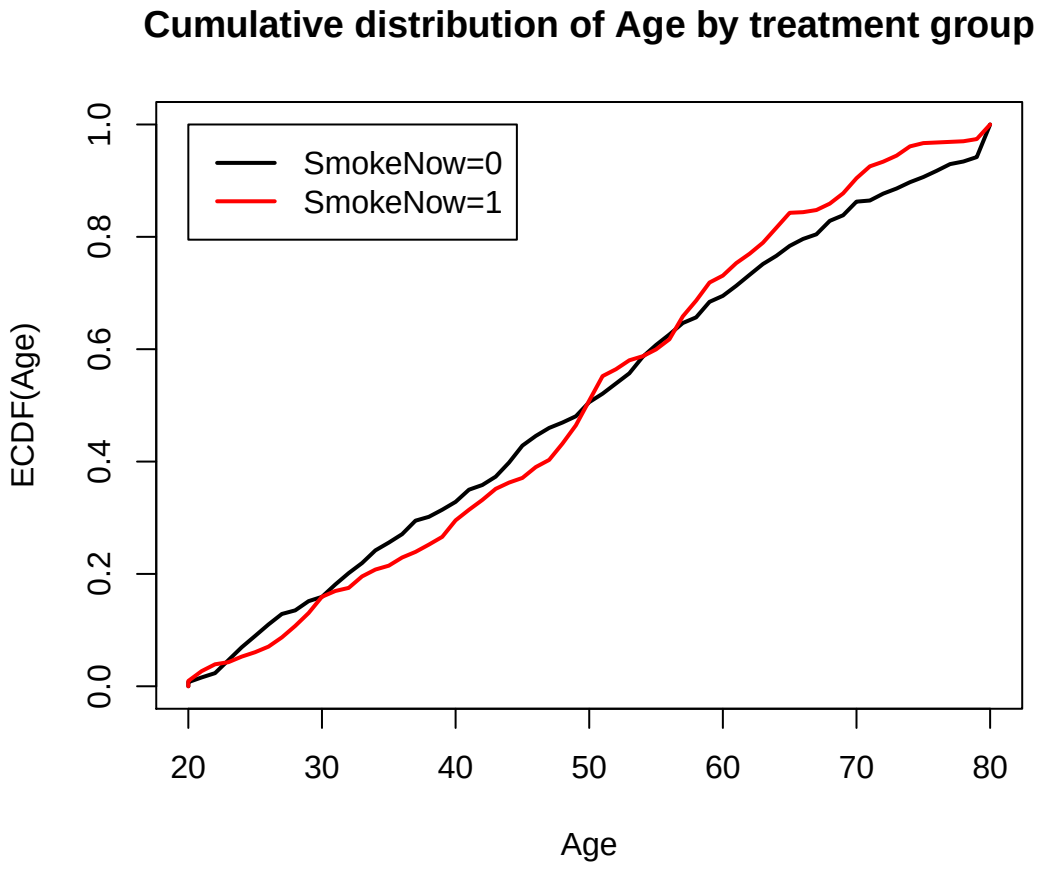



```
rbind(table(cut(ps.lr[Smoke==0],dec)),table(cut(ps.lr[Smoke==1],dec)))

+      (0,0.176] (0.176,0.222] (0.222,0.286] (0.286,0.34] (0.34,0.404]
+ [1,]       118           113           102           92           87
+ [2,]        20            27            35           47           48
+      (0.404,0.481] (0.481,0.564] (0.564,0.639] (0.639,0.725] (0.725,0.941]
+ [1,]            80            66            55            37            32
+ [2,]            57            74            83           100           104

ps.lr.dec <- cut(ps.lr,dec,labels=1:10)
SMD.10.table <- ExtractSmd(tabUnmatched)
for(j in 1:10) {
        tabPSdec <- CreateTableOne(vars = vars, strata = "SmokeNow",
                    data = small.nhanes[ps.lr.dec==j,], test = FALSE)
        SMD.10.table <- cbind(SMD.10.table,ExtractSmd(tabPSdec))
}
round(SMD.10.table,2)

+               SMD.10.table
+ Gender                0.14 0.07 0.29 0.23 0.12 0.21 0.23 0.20 0.25 0.06
+ Age                   0.59 0.41 0.12 0.14 0.12 0.14 0.01 0.02 0.64 0.16
+ Race3                 0.31 0.46 0.51 0.31 0.10 0.41 0.43 0.62 0.34 0.38
+ Education             0.51 0.49 0.63 0.76 0.40 0.51 0.38 0.31 0.61 0.34
+ MaritalStatus         0.49 0.69 0.50 0.78 0.43 0.49 0.17 0.27 0.43 0.38
+ Poverty               0.45 0.03 0.08 0.10 0.13 0.09 0.13 0.03 0.02 0.07
+
+ Gender         0.09
+ Age            0.12
```

```
+ Race3          0.69
+ Education     0.54
+ MaritalStatus 0.32
+ Poverty       0.35
```

3. Inverse Probability Weighting

```
ps.lr.weight <-Smoke/ps.lr + (1-Smoke)/(1-ps.lr)
temp0 <- Ecdf(small.nhanes$Age[Smoke==0],weights=ps.lr.weight[Smoke==0],pl=F)
temp1 <- Ecdf(small.nhanes$Age[Smoke==1],weights=ps.lr.weight[Smoke==1],pl=F)
plot(temp0$x,temp0$y,ylab="ECDF(Age)",xlab="Age",
    main="Cumulative distribution of Age by treatment group",type="l",lwd=2)
lines(temp1$x,temp1$y,col="red",lwd=2)
legend(20,1,c('SmokeNow=0','SmokeNow=1'),col=c('black','red'),lty=1,lwd=2)
```




```
nhanes.IPW.lr <- svydesign(ids=~0, data=small.nhanes, weights=ps.lr.weight)
tabIPW <- svyCreateTableOne(vars = vars, strata = "SmokeNow", data = nhanes.IPW.lr, test = FALSE)
print(tabIPW, smd = TRUE)

+                      Stratified by SmokeNow
+                       0               1               SMD
+   n                     1379.0          1379.7
+   Gender = male (%)      797.4 (57.8)    781.8 (56.7)   0.023
```

```
+   Age (mean (sd))       50.06 (17.16)   49.83 (15.42)  0.014
+   Race3 (%)                                            0.052
+      Asian               40.4 ( 2.9)     37.9 ( 2.7)
+      Black              119.0 ( 8.6)    103.3 ( 7.5)
+      Hispanic            75.5 ( 5.5)     70.6 ( 5.1)
+      Mexican             81.2 ( 5.9)     87.9 ( 6.4)
+      White             1019.8 (74.0)   1038.8 (75.3)
+      Other               43.0 ( 3.1)     41.2 ( 3.0)
+   Education (%)                                        0.029
+      8th Grade           89.4 ( 6.5)     96.2 ( 7.0)
+      9 - 11th Grade     184.6 (13.4)    185.8 (13.5)
+      High School        293.9 (21.3)    285.5 (20.7)
+      Some College       482.7 (35.0)    474.2 (34.4)
+      College Grad       328.4 (23.8)    338.1 (24.5)
+   MaritalStatus (%)                                    0.023
+      Divorced           161.0 (11.7)    165.1 (12.0)
+      LivePartner        154.6 (11.2)    150.8 (10.9)
+      Married            695.0 (50.4)    690.1 (50.0)
+      NeverMarried       249.9 (18.1)    254.0 (18.4)
+      Separated           23.1 ( 1.7)     20.5 ( 1.5)
+      Widowed             95.4 ( 6.9)     99.2 ( 7.2)
+   Poverty (mean (sd))    2.80 (1.67)     2.80 (1.63)  <0.001
```

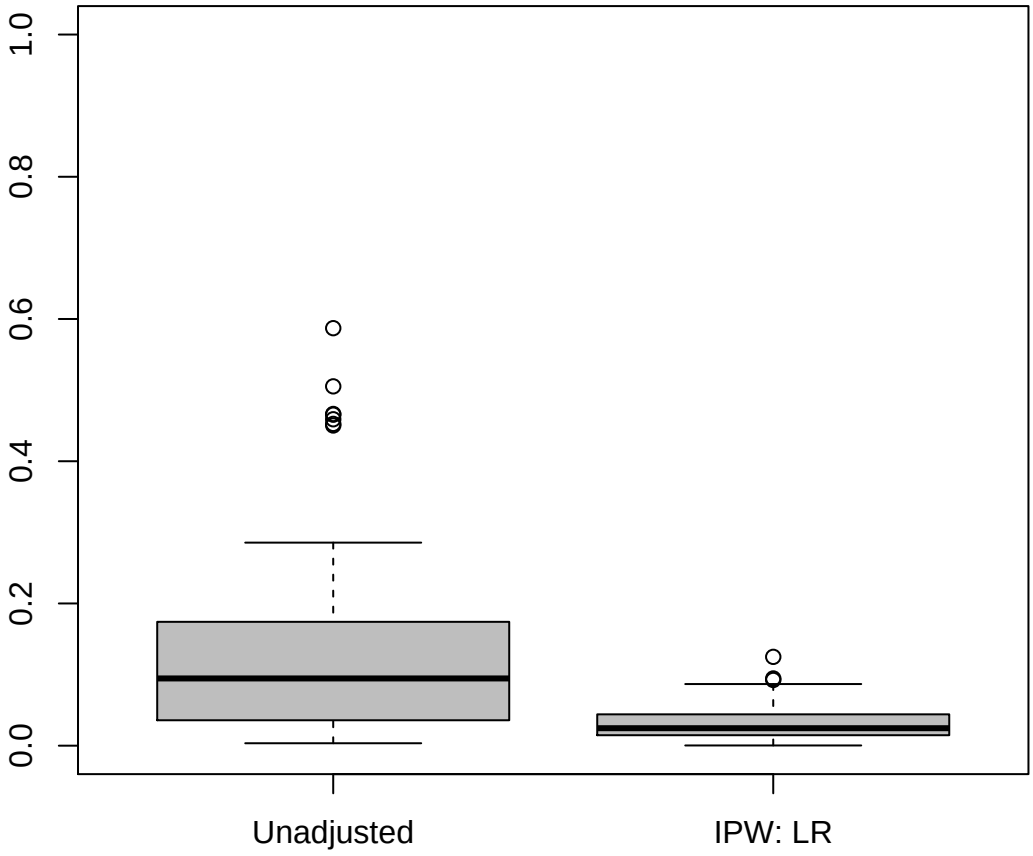
```
round(cbind(SMD.table,ExtractSmd(tabIPW)),3)
```

```
+               SMD.table
+ Gender            0.138 0.102 0.104 0.029 0.200 0.031 0.023
+ Age               0.592 0.257 0.171 0.099 0.311 0.164 0.014
+ Race3             0.315 0.317 0.112 0.344 0.415 0.287 0.052
+ Education         0.512 0.538 0.417 0.280 0.238 0.302 0.029
+ MaritalStatus     0.488 0.432 0.239 0.272 0.233 0.261 0.023
+ Poverty           0.453 0.087 0.126 0.114 0.004 0.146 0.000
```

```
interact.data$ps.lr.weight <- ps.lr.weight
nhanes.IPW.lr.interact <- svydesign(ids=~0, data=interact.data, weights=ps.lr.weight)
SMDinteract.IPW.a <- svyCreateTableOne(vars = vars.interact[1:25],
       strata = "SmokeNow", data = nhanes.IPW.lr.interact, test = FALSE)
SMDinteract.IPW.b <- svyCreateTableOne(vars = vars.interact[26:50],
       strata = "SmokeNow", data = nhanes.IPW.lr.interact, test = FALSE)
SMDinteract.IPW.c <- svyCreateTableOne(vars = vars.interact[51:78],
       strata = "SmokeNow", data = nhanes.IPW.lr.interact, test = FALSE)
SMDinteract.IPW <- c(ExtractSmd(SMDinteract.IPW.a),ExtractSmd(SMDinteract.IPW.b),
                     ExtractSmd(SMDinteract.IPW.c))
summary(SMDinteract.IPW)
```

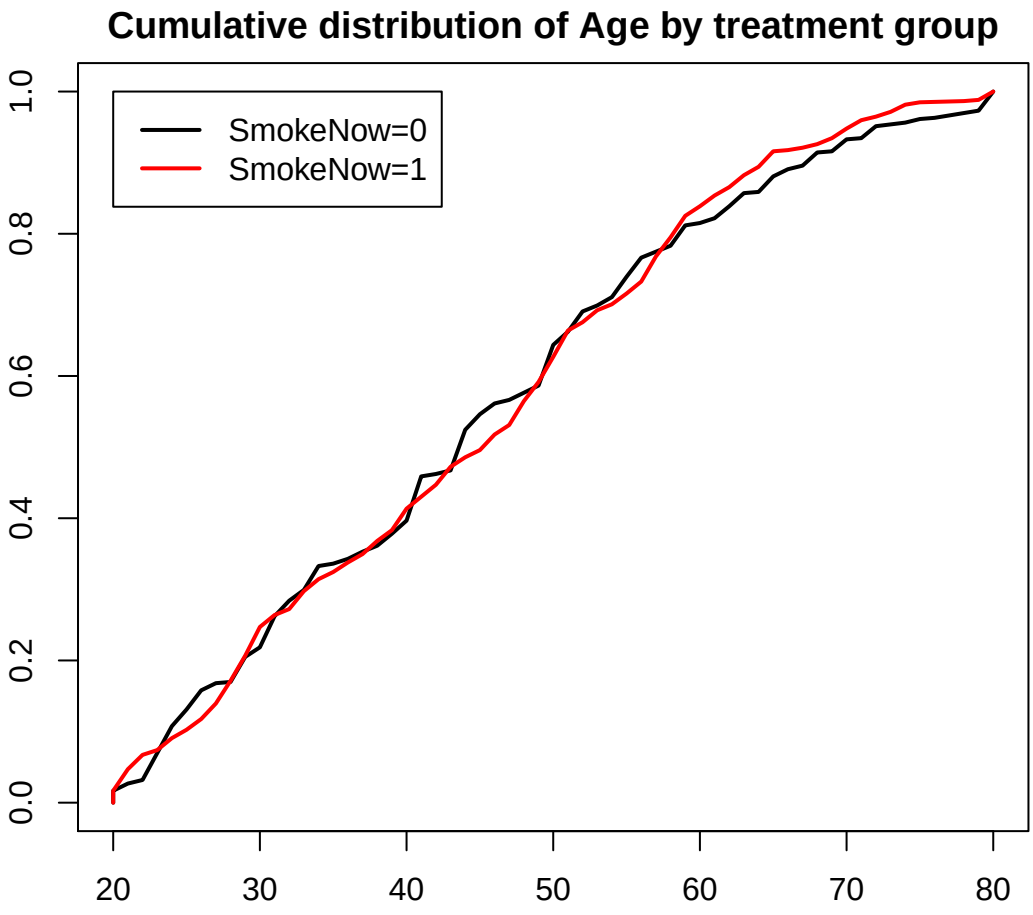
```
+      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
+ 0.0002997 0.0147610 0.0245889 0.0314200 0.0439615 0.1249524
```

```
interact.table <- cbind(ExtractSmd(SMDinteract),SMDinteract.IPW)
par(mar=c(3,3,2,1))
boxplot(cbind(ExtractSmd(SMDinteract),SMDinteract.IPW),col='gray',
               ylim=range(0,1),names=c("Unadjusted","IPW: LR"))
```



4. Matching

```
## Matching
ps.lr.match <- Match(Tr=small.nhanes$SmokeNow,X=small.nhanes$ps.lr,ties=FALSE)
matched.samp <- small.nhanes[c(ps.lr.match$index.control,ps.lr.match$index.treated),]
table(table(c(ps.lr.match$index.control,ps.lr.match$index.treated)))

+
+   1   2   3   4   5   6   7   8   9  11  12  13  14  15
+ 725  60  25  21   9   6   2   1   2   1   1   1   1   1

temp0 <- Ecdf(matched.samp$Age[matched.samp$SmokeNow==0],pl=F)
temp1 <- Ecdf(matched.samp$Age[matched.samp$SmokeNow==1],pl=F)
par(mar=c(3,3,2,1))
plot(temp0$x,temp0$y,ylab="ECDF(Age)",xlab="Age",main="Cumulative distribution of Age by treatment group",t
lines(temp1$x,temp1$y,col="red",lwd=2)
legend(20,1,c('SmokeNow=0','SmokeNow=1'),col=c('black','red'),lwd=2)
```

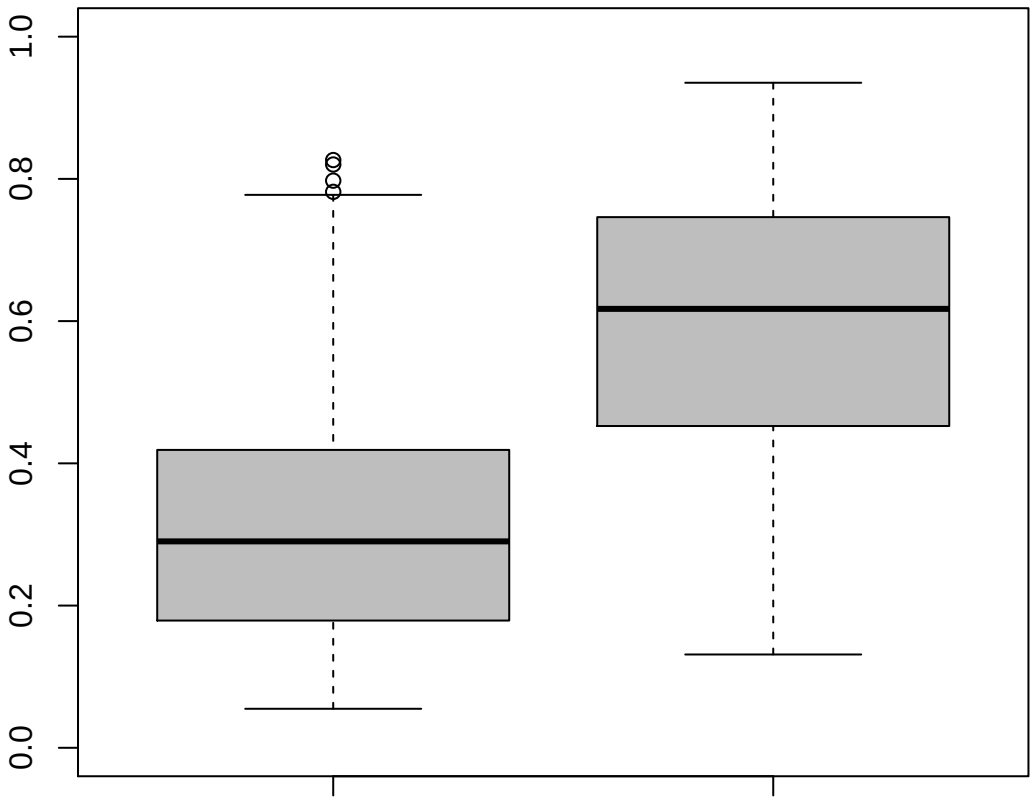



```
##Summarize balance after matching: ##Not run
#MatchBalance(SmokeNow~Gender+Age+Race3+Education+MaritalStatus+HHIncome+Poverty,
#                   data=small.nhanes,match.out=ps.lr.match)

tabMatched <- CreateTableOne(vars = vars, strata = "SmokeNow", data = matched.samp, test = FALSE)
round(cbind(SMD.table,ExtractSmd(tabMatched),ExtractSmd(tabIPW)),3)

+               SMD.table
+ Gender            0.138 0.102 0.104 0.029 0.200 0.031 0.120 0.023
+ Age               0.592 0.257 0.171 0.099 0.311 0.164 0.010 0.014
+ Race3             0.315 0.317 0.112 0.344 0.415 0.287 0.146 0.052
+ Education         0.512 0.538 0.417 0.280 0.238 0.302 0.109 0.029
+ MaritalStatus     0.488 0.432 0.239 0.272 0.233 0.261 0.185 0.023
+ Poverty           0.453 0.087 0.126 0.114 0.004 0.146 0.096 0.000

Max.SMD <- c(NA,Max.SMD,max(ExtractSmd(tabMatched)),NA,max(ExtractSmd(tabIPW)))
Mean.SMD <- c(NA,Mean.SMD,mean(ExtractSmd(tabMatched)),NA,mean(ExtractSmd(tabIPW)))
Med.SMD <- c(NA,Med.SMD,median(ExtractSmd(tabMatched)),NA,median(ExtractSmd(tabIPW)))
ps.SMDtable <- round(cbind(SMD.table,ExtractSmd(tabMatched),ExtractSmd(tabIPW)),3)
```

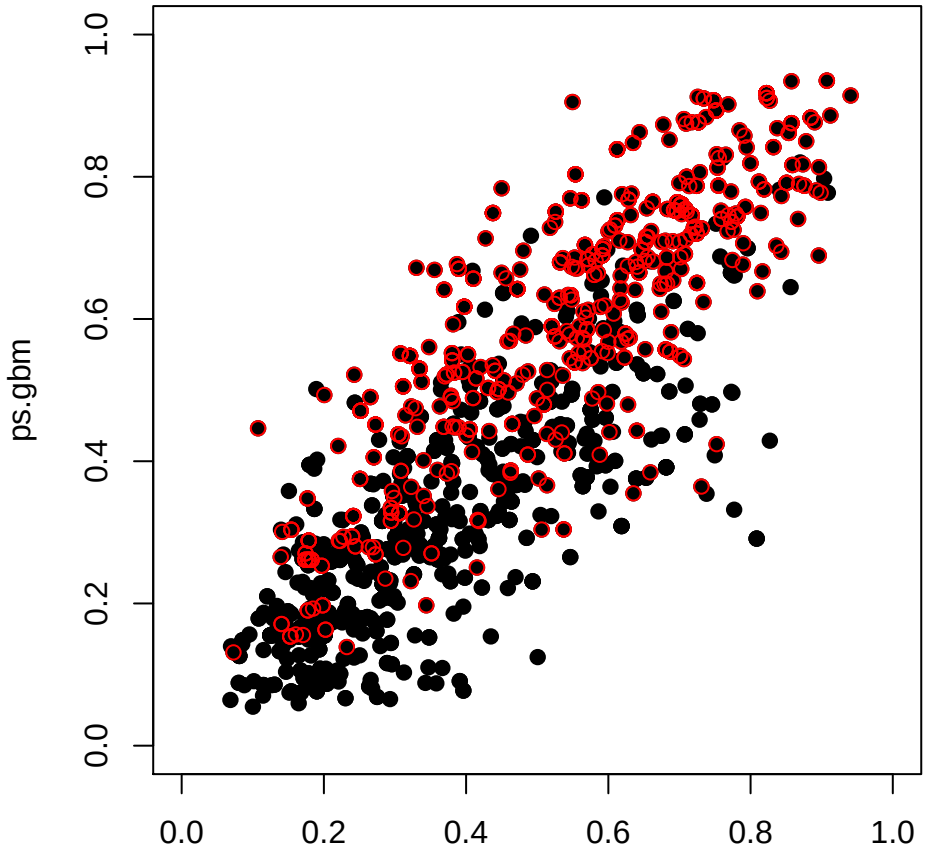
5. Generalized Boosting

```
## GBM with twang library
gbm.fit <- ps(SmokeNow~Gender+Age+Race3+Education+MaritalStatus+HHIncome+Poverty, estimand = "ATE",
```

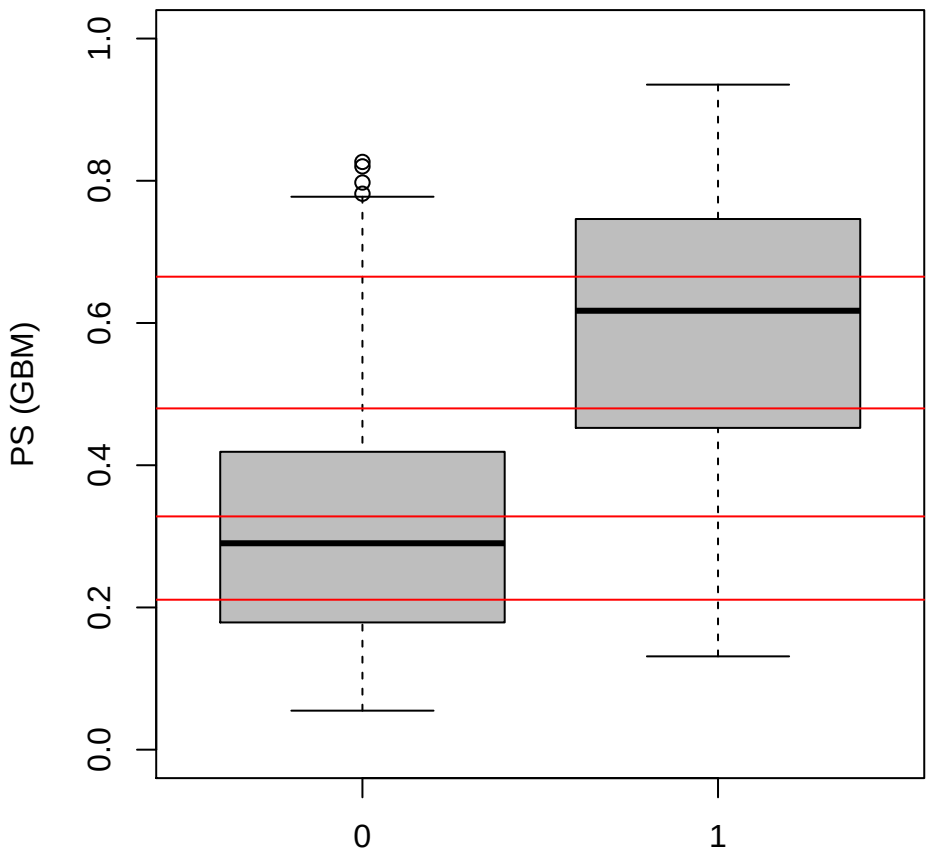
```
                data=as.data.frame(small.nhanes),verbose=FALSE)
ps.gbm <- gbm.fit$ps$ks.mean.ATE;small.nhanes$ps.gbm <- ps.gbm
summary(ps.gbm)

+    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
+ 0.05486 0.25629 0.40941 0.43210 0.60759 0.93518

par(mar=c(3,3,2,1))
boxplot(ps.gbm[small.nhanes$SmokeNow==0],ps.gbm[small.nhanes$SmokeNow==1],col='gray',ylim=range(0,1))
```




```
par(mar=c(3,3,2,1),pty='s')
plot(ps.lr,ps.gbm,pch=19,xlim=range(0,1),ylim=range(0,1))  ## correlate pretty well
points(ps.lr[small.nhanes$SmokeNow==1],ps.gbm[small.nhanes$SmokeNow==1],col="red")  ## correlate very badly
```

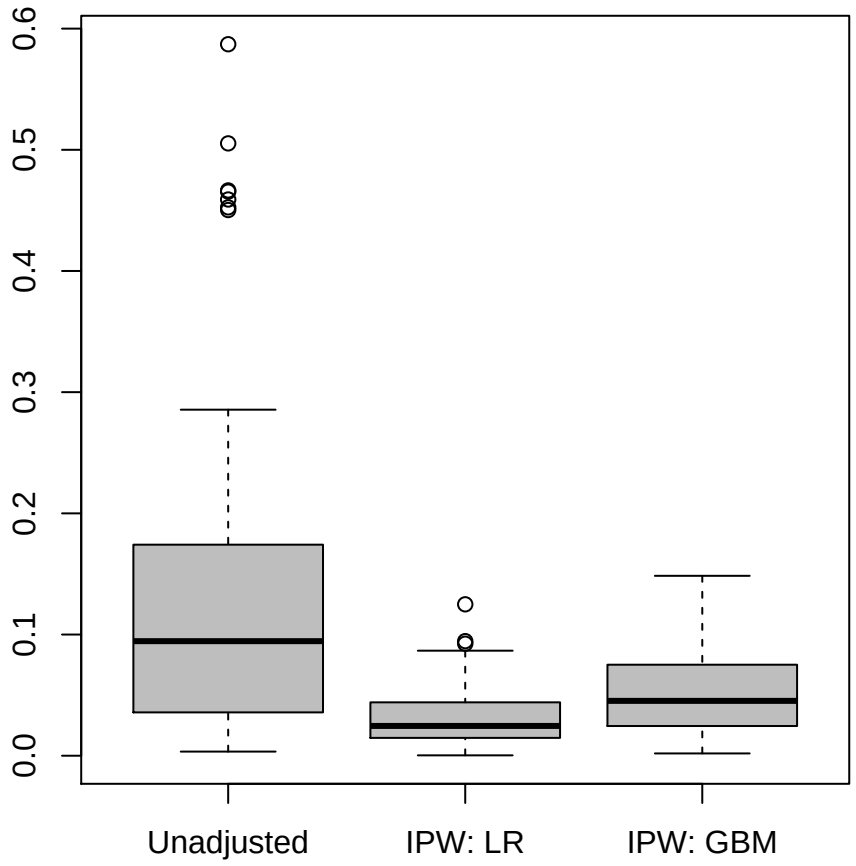



```
par(mar=c(3,3,2,1))
boxplot(ps.gbm[small.nhanes$SmokeNow==0],ps.gbm[small.nhanes$SmokeNow==1],ylim=range(0,1),
             ylab="PS (GBM)",xlab="Treatment Group",names=c(0,1),col='gray')
(quints.gbm <- c(0,quantile(ps.gbm,seq(.2,1,.2))))

+                 20%       40%       60%       80%      100%
+ 0.0000000 0.2108766 0.3280177 0.4799049 0.6651142 0.9351823

abline(h=quints.gbm[2:5],col="red")
```



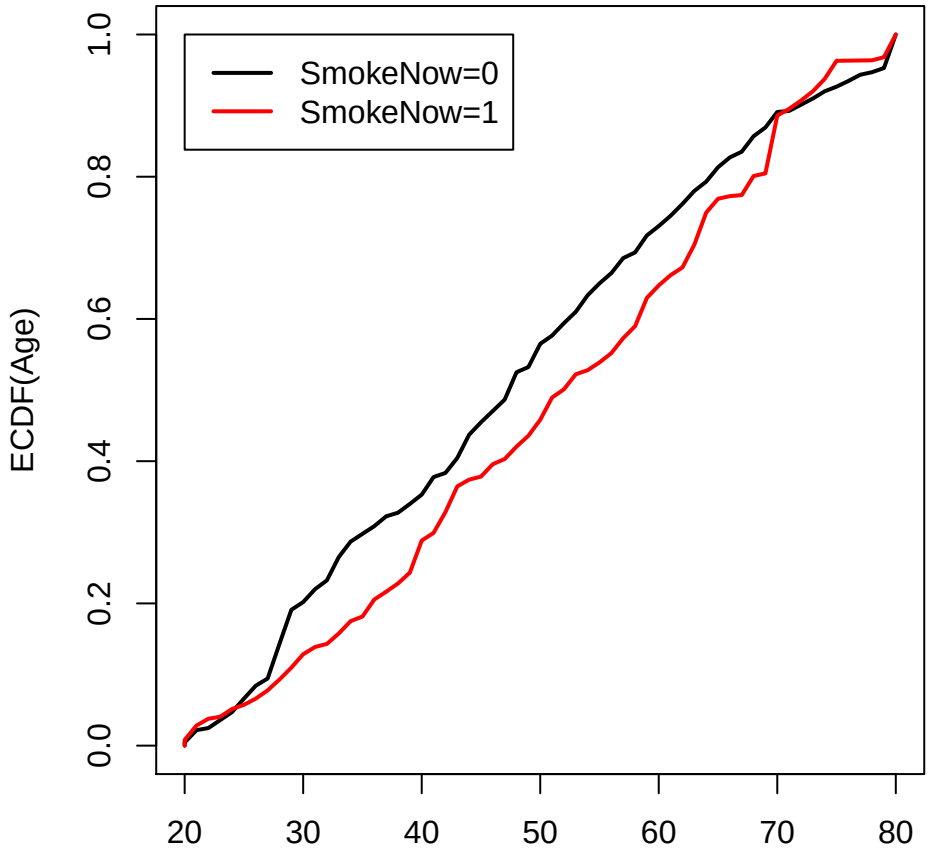

```
small.nhanes$ps.gbm <- ps.gbm

rbind(table(cut(ps.gbm[small.nhanes$SmokeNow==0],quints.gbm)),
table(cut(ps.gbm[small.nhanes$SmokeNow==1],quints.gbm)))

+      (0,0.211] (0.211,0.328] (0.328,0.48] (0.48,0.665] (0.665,0.935]
+ [1,]       259           209          195          100            19
+ [2,]        17            66           81          175           256

ps.gbm.weight <- small.nhanes$SmokeNow/ps.gbm + (1-small.nhanes$SmokeNow)/(1-ps.gbm)

nhanes.IPW.gbm <- svydesign(ids=~0, data=small.nhanes, weights=ps.gbm.weight)
tabIPW.gbm <- svyCreateTableOne(vars = vars, strata = "SmokeNow", data = nhanes.IPW.gbm, test = FALSE)
round(cbind(ExtractSmd(tabUnmatched),ExtractSmd(tabIPW),ExtractSmd(tabIPW.gbm)),3)

+                [,1]  [,2]  [,3]
+ Gender         0.138 0.023 0.065
+ Age            0.592 0.014 0.168
+ Race3          0.315 0.052 0.116
+ Education      0.512 0.029 0.153
+ MaritalStatus 0.488 0.023 0.156
+ Poverty        0.453 0.000 0.096

##########
## on interactions?
interact.data$ps.gbm.weight <- ps.gbm.weight
nhanes.IPW.gbm.interact <- svydesign(ids=~0, data=interact.data, weights=ps.gbm.weight)
SMDinteract.IPW.gbm.a <- svyCreateTableOne(vars = vars.interact[1:25],
```

```
       strata = "SmokeNow", data = nhanes.IPW.gbm.interact, test = FALSE)
SMDinteract.IPW.gbm.b <- svyCreateTableOne(vars = vars.interact[26:50],
       strata = "SmokeNow", data = nhanes.IPW.gbm.interact, test = FALSE)
SMDinteract.IPW.gbm.c <- svyCreateTableOne(vars = vars.interact[51:78],
       strata = "SmokeNow", data = nhanes.IPW.gbm.interact, test = FALSE)
SMDinteract.IPW.gbm <- c(ExtractSmd(SMDinteract.IPW.gbm.a),ExtractSmd(SMDinteract.IPW.gbm.b),
                         ExtractSmd(SMDinteract.IPW.gbm.c))
summary(SMDinteract.IPW.gbm)

+     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
+ 0.001895 0.025358 0.045265 0.053004 0.074237 0.148453

interact.table <- cbind(interact.table,SMDinteract.IPW.gbm)
boxplot(interact.table,col='gray',names=c("Unadjusted","IPW: LR", "IPW: GBM"))
```

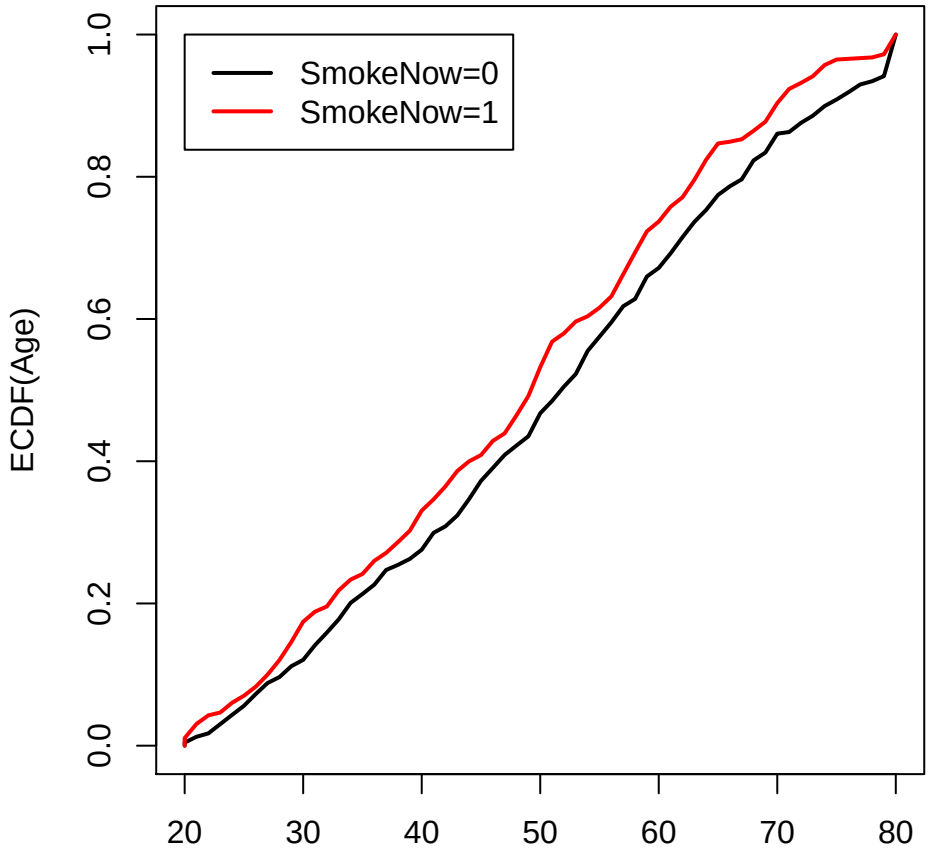



```
ps.gbm.quints <- cut(ps.gbm,quints.gbm,labels=1:5)
SMD.gbm.table <- NULL
for(j in 1:5) {
  tabPSquints.gbm <- CreateTableOne(vars = vars, strata = "SmokeNow",
                           data = small.nhanes[ps.gbm.quints==j,], test = FALSE)
  SMD.gbm.table <- cbind(SMD.gbm.table,ExtractSmd(tabPSquints.gbm))
}
round(SMD.gbm.table,3)

+                [,1]  [,2]  [,3]  [,4]  [,5]
+ Gender        0.040 0.168 0.065 0.279 0.109
+ Age           0.027 0.450 0.273 0.272 0.142
```

```
+ Race3          0.512 0.434 0.163 0.310 0.693
+ Education     0.576 0.267 0.368 0.438 0.702
+ MaritalStatus 0.716 0.682 0.665 0.318 0.499
+ Poverty       0.372 0.118 0.205 0.542 0.099

ps.gbm.match <- Match(Tr=small.nhanes$SmokeNow,X=small.nhanes$ps.gbm,estimand="ATE",ties=FALSE)
matched.gbm.samp <- small.nhanes[c(ps.gbm.match$index.control,ps.gbm.match$index.treated),]

#Not run
#MatchBalance(SmokeNow~Gender+Age+Race3+Education+MaritalStatus+HHIncome+Poverty,
#                    data=small.nhanes,match.out=ps.gbm.match)
tabMatched.gbm <- CreateTableOne(vars = vars, strata = "SmokeNow", data = matched.gbm.samp, test = FALSE)


temp0 <- Ecdf(matched.gbm.samp$Age[matched.gbm.samp$SmokeNow==0],pl=F)
temp1 <- Ecdf(matched.gbm.samp$Age[matched.gbm.samp$SmokeNow==1],pl=F)
plot(temp0$x,temp0$y,ylab="ECDF(Age)",xlab="Age",main="",type="l",lwd=2)
lines(temp1$x,temp1$y,col="red",lwd=2)
legend(20,1,c('SmokeNow=0','SmokeNow=1'),col=c('black','red'),lty=1,lwd=2)
```




```
temp0<-Ecdf(small.nhanes$Age[Smoke==0],weights=ps.gbm.weight[Smoke==0],pl=F)
temp1<-Ecdf(small.nhanes$Age[Smoke==1],weights=ps.gbm.weight[Smoke==1],pl=F)
plot(temp0$x,temp0$y,ylab="ECDF(Age)",xlab="Age",main="",type="l",lwd=2)
lines(temp1$x,temp1$y,col="red",lwd=2)
legend(20,1,c('SmokeNow=0','SmokeNow=1'),col=c('black','red'),lty=1,lwd=2)
```

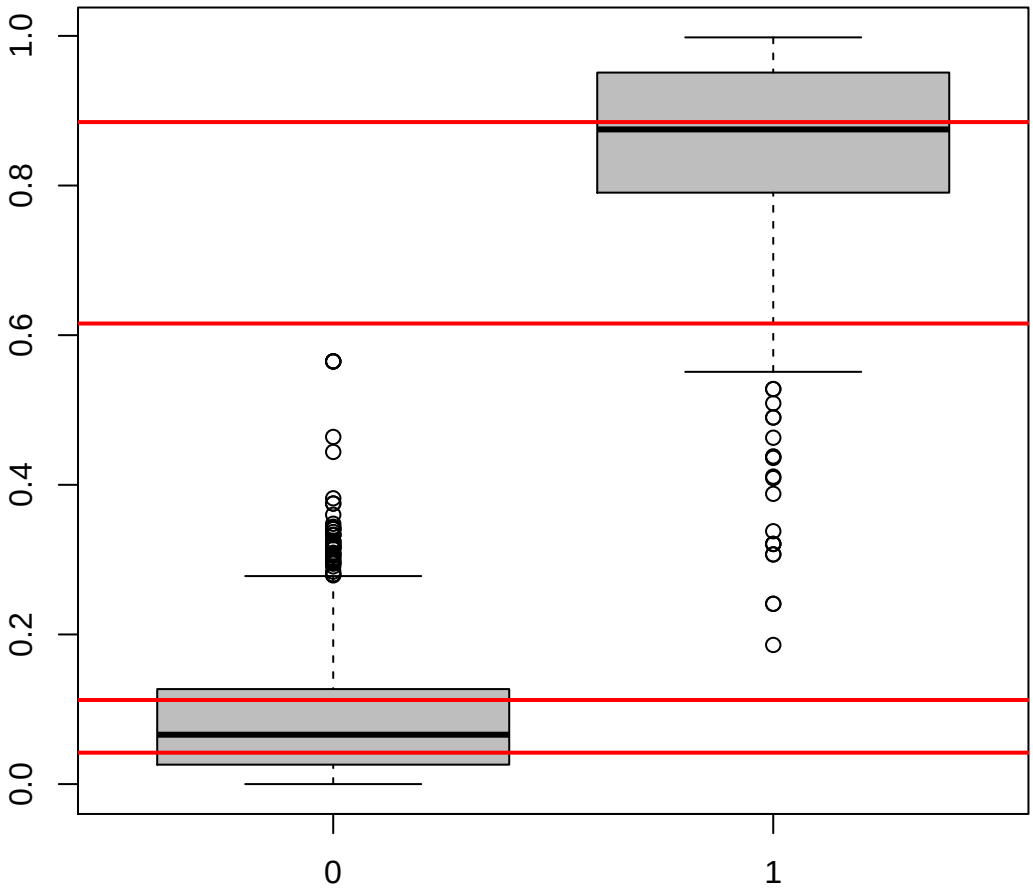



```
SMD.gbm.table <- cbind(SMD.gbm.table,ExtractSmd(tabMatched.gbm),ExtractSmd(tabIPW.gbm))
round(SMD.gbm.table,3)

+                [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]
+ Gender         0.040 0.168 0.065 0.279 0.109 0.095 0.065
+ Age            0.027 0.450 0.273 0.272 0.142 0.205 0.168
+ Race3          0.512 0.434 0.163 0.310 0.693 0.291 0.116
+ Education      0.576 0.267 0.368 0.438 0.702 0.284 0.153
+ MaritalStatus 0.716 0.682 0.665 0.318 0.499 0.145 0.156
+ Poverty        0.372 0.118 0.205 0.542 0.099 0.237 0.096
```

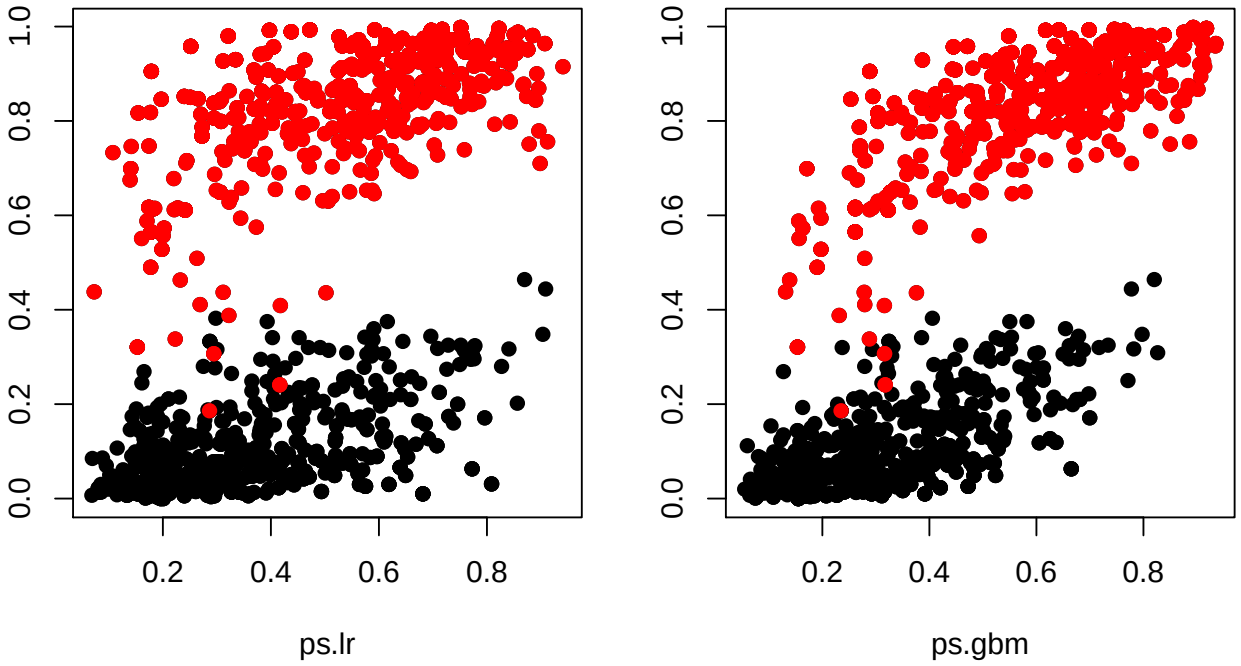
6. Super Learner

```
## SuperLearner
X.mat <- data.frame(cbind(small.nhanes$Gender,small.nhanes$Age,small.nhanes$Race3,
    small.nhanes$Education,small.nhanes$MaritalStatus,small.nhanes$HHIncome,small.nhanes$Poverty))
my.library <- c("SL.knn","SL.randomForest","SL.glmnet","SL.mean")
SL.fit <- SuperLearner(Y = small.nhanes$SmokeNow, X = X.mat,
                  SL.library = my.library,verbose = FALSE, method ="method.NNLS",
                  family=binomial())
small.nhanes$ps.SL <- ps.SL <- SL.fit$SL.predict
mean(as.numeric(ps.SL > 0.5) == Smoke)

+ [1] 0.9840232
```

```
par(mar=c(2,3,2,1))
boxplot(ps.SL[Smoke==0],ps.SL[Smoke==1],ylab="PS (SL)",xlab="Treatment Group",names=c(0,1),col='gray')
(quints.SL <- c(0,quantile(ps.SL,seq(.2,1,.2))))

+           20%    40%    60%    80%   100%
+ 0.0000 0.0420 0.1124 0.6156 0.8848 0.9980

abline(h=quints.SL[2:5],col="red",lwd=2)
```



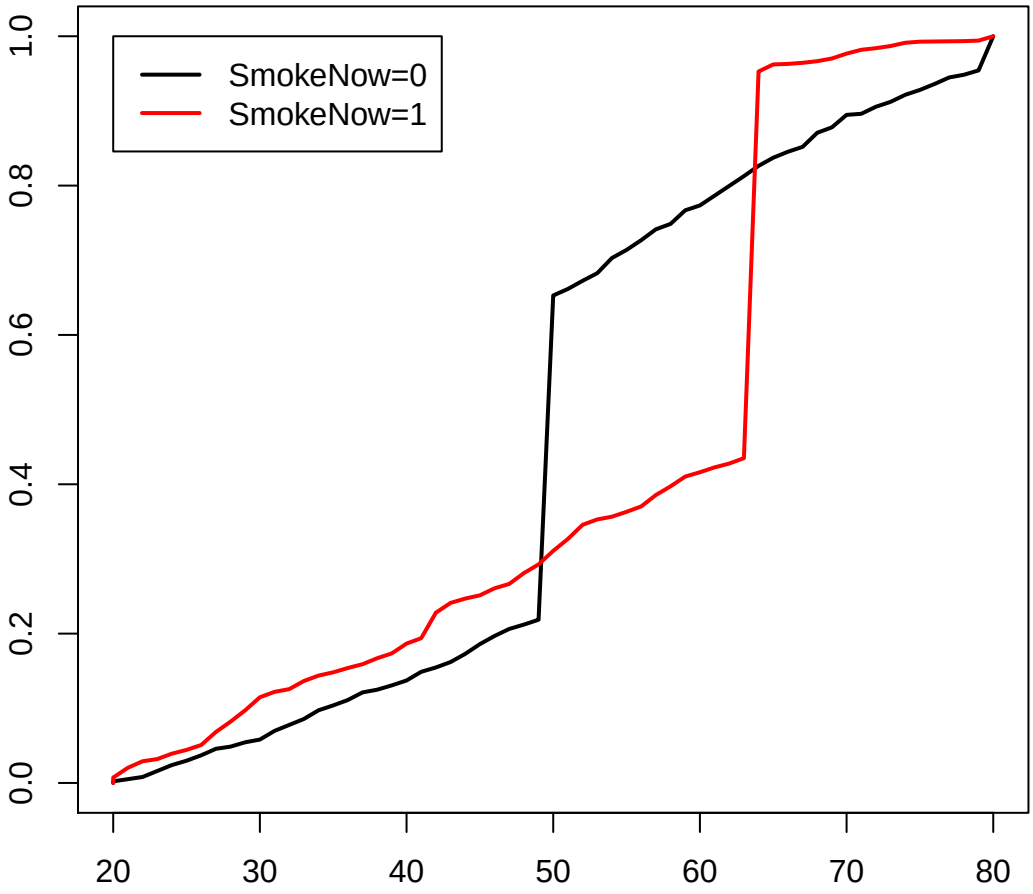

```
par(mar=c(2,3,2,1),mfrow=c(1,2),pty='s')
plot(ps.lr,ps.SL,pch=19)  ## correlate very badly!!
points(ps.lr[Smoke==1],ps.SL[Smoke==1],col="red",pch=19)  ## correlate very badly!!
plot(ps.gbm,ps.SL,pch=19)  ## correlate badly
points(ps.gbm[Smoke==1],ps.SL[Smoke==1],col="red",pch=19)  ## correlate very badly!!
```




```
rbind(table(cut(ps.SL[Smoke==0],quints.SL)),
table(cut(ps.SL[Smoke==1],quints.SL)))

+      (0,0.042] (0.042,0.112] (0.112,0.616] (0.616,0.885] (0.885,0.998]
+ [1,]       268           274           231             0             0
+ [2,]         0             0            44           275           276

ps.SL[ps.SL==0] <- min(ps.SL[ps.SL!=0])
ps.SL.weight <- Smoke/ps.SL + (1-Smoke)/(1-ps.SL)
nhanes.IPW.SL <- svydesign(ids=~0, data=small.nhanes, weights=ps.SL.weight)
tabIPW.SL <- svyCreateTableOne(vars = vars, strata = "SmokeNow",
                               data = nhanes.IPW.SL, test = FALSE)
ExtractSmd(tabIPW.SL)

+        Gender           Age         Race3     Education MaritalStatus
+     0.1135016     0.4900814     0.2478273     0.4073276     0.3833316
+       Poverty
+     0.2958973

ps.SL.quints <- cut(ps.SL,quints.SL,labels=1:5)
SMD.SL.table <- NULL
for(j in 3) {
  tabPSquints.SL <- CreateTableOne(vars = vars, strata = "SmokeNow",
                                   data = small.nhanes[ps.SL.quints==j,], test = FALSE)
  SMD.SL.table <- cbind(SMD.SL.table,ExtractSmd(tabPSquints.SL))
}
round(SMD.SL.table,3)
```

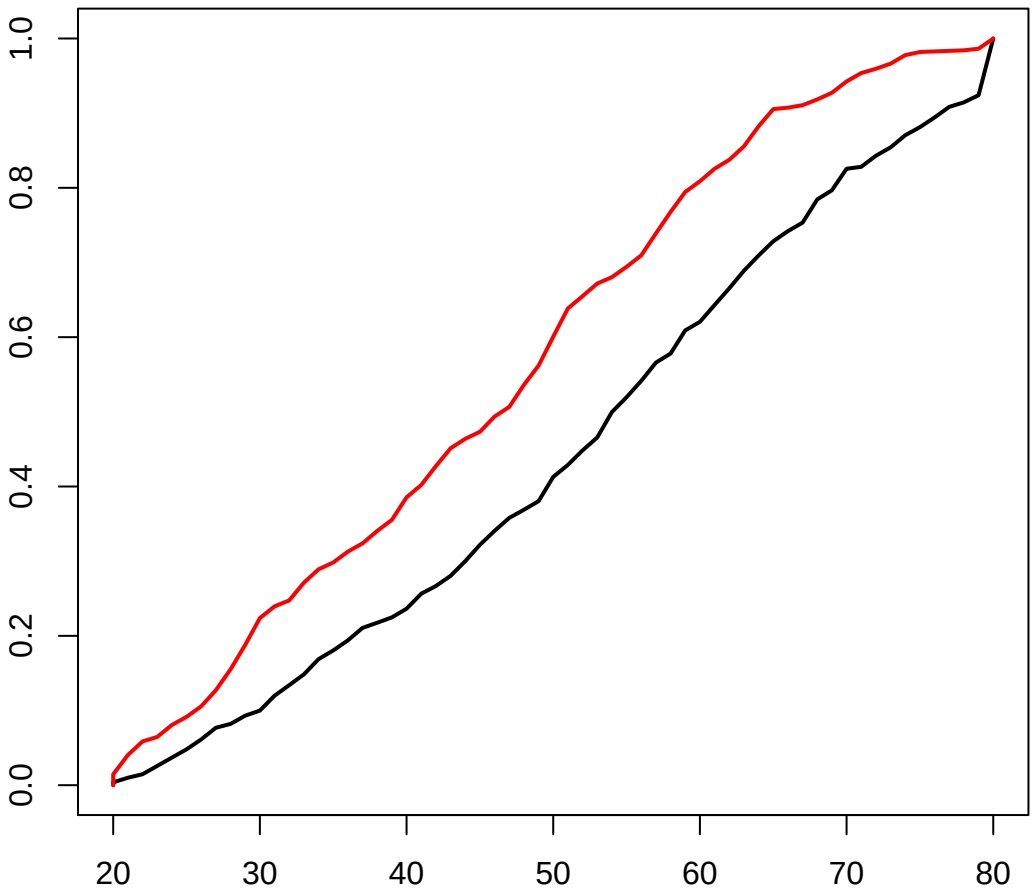
```
+                 [,1]
+ Gender         0.460
+ Age            0.582
+ Race3          1.023
+ Education      1.002
+ MaritalStatus 0.740
+ Poverty        1.585
```

```
ps.SL.match <- Match(Tr=small.nhanes$SmokeNow,X=small.nhanes$ps.SL,estimand="ATE",ties=FALSE)
matched.SL.samp <- small.nhanes[c(ps.SL.match$index.control,ps.SL.match$index.treated),]

##Not run
#MatchBalance(SmokeNow~Gender+Age+Race3+Education+MaritalStatus+HHIncome+Poverty,
#                    data=small.nhanes,match.out=ps.SL.match)
tabMatched.SL <- CreateTableOne(vars = vars, strata = "SmokeNow",data = matched.SL.samp, test = FALSE)

temp0 <- Ecdf(matched.SL.samp$Age[matched.SL.samp$SmokeNow==0],pl=F)
temp1 <- Ecdf(matched.SL.samp$Age[matched.SL.samp$SmokeNow==1],pl=F)
par(mar=c(2,3,2,1))
plot(temp0$x,temp0$y,ylab="ECDF(Age)",xlab="Age",main="",type="l",lwd=2)
lines(temp1$x,temp1$y,col="red",lwd=2)
legend(20,1,c('SmokeNow=0','SmokeNow=1'),col=c('black','red'),lty=1,lwd=2)
```



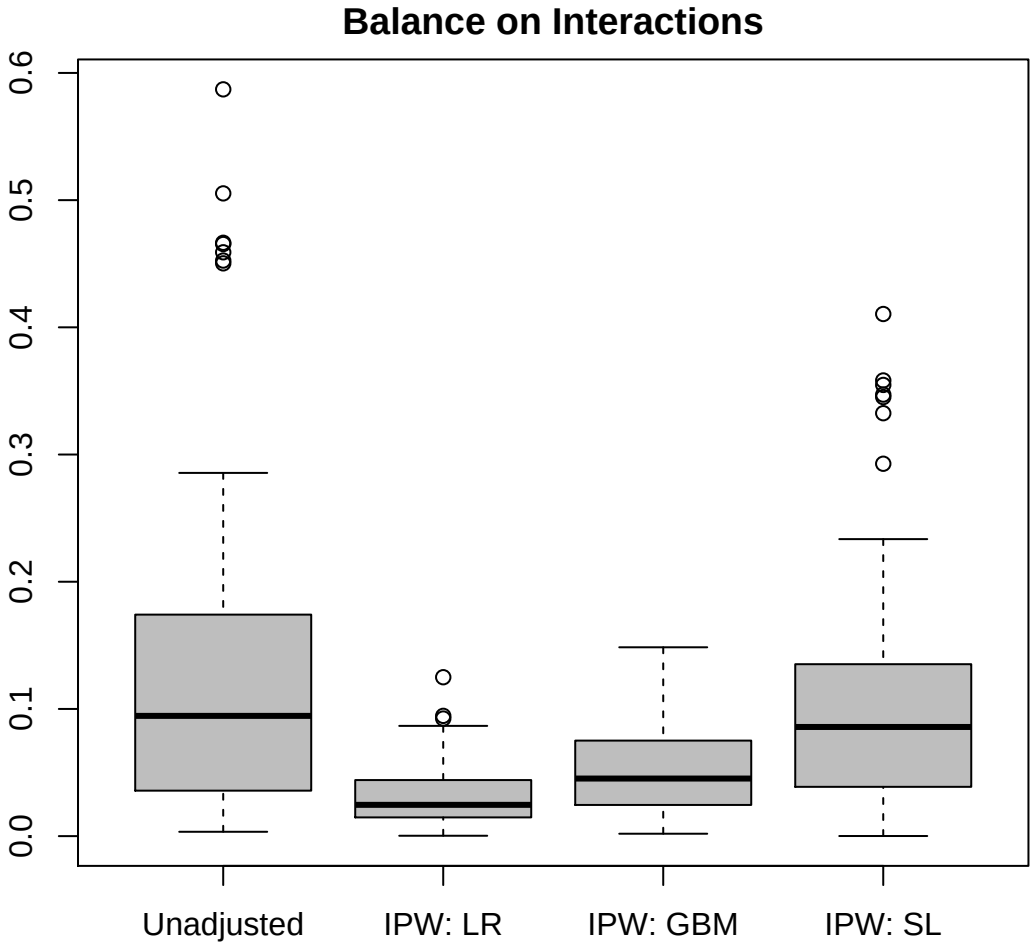

```
temp0 <- Ecdf(small.nhanes$Age[Smoke==0],weights=ps.SL.weight[Smoke==0],pl=F)
temp1 <- Ecdf(small.nhanes$Age[Smoke==1],weights=ps.SL.weight[Smoke==1],pl=F)
par(mar=c(2,3,2,1))
```

```
plot(temp0$x,temp0$y,ylab="ECDF(Age)",xlab="Age",main="",type="l",lwd=2)
lines(temp1$x,temp1$y,col="red",lwd=2)
```




```
SMD.SL.table <- cbind(SMD.SL.table,ExtractSmd(tabMatched.SL),ExtractSmd(tabIPW.SL))
round(SMD.SL.table,3)

+                [,1]  [,2]  [,3]
+ Gender         0.460 1.128 0.114
+ Age            0.582 0.170 0.490
+ Race3          1.023 0.138 0.248
+ Education      1.002 1.538 0.407
+ MaritalStatus 0.740 0.231 0.383
+ Poverty        1.585 0.006 0.296

round(cbind(ExtractSmd(tabUnmatched),ExtractSmd(tabIPW),
            ExtractSmd(tabIPW.gbm),ExtractSmd(tabIPW.SL)),3)

+                [,1]  [,2]  [,3]  [,4]
+ Gender         0.138 0.023 0.065 0.114
+ Age            0.592 0.014 0.168 0.490
+ Race3          0.315 0.052 0.116 0.248
+ Education      0.512 0.029 0.153 0.407
+ MaritalStatus 0.488 0.023 0.156 0.383
+ Poverty        0.453 0.000 0.096 0.296

##########
## Balance on interactions?
```

```
interact.data$ps.SL.weight <- ps.SL.weight
nhanes.IPW.SL.interact <- svydesign(ids=~0, data=interact.data, weights=ps.SL.weight)
SMDinteract.IPW.SL.a <- svyCreateTableOne(vars = vars.interact[1:25],
       strata = "SmokeNow", data = nhanes.IPW.SL.interact, test = FALSE)
SMDinteract.IPW.SL.b <- svyCreateTableOne(vars = vars.interact[26:50],
       strata = "SmokeNow", data = nhanes.IPW.SL.interact, test = FALSE)
SMDinteract.IPW.SL.c <- svyCreateTableOne(vars = vars.interact[51:78],
       strata = "SmokeNow", data = nhanes.IPW.SL.interact, test = FALSE)
SMDinteract.IPW.SL <- c(ExtractSmd(SMDinteract.IPW.SL.a),ExtractSmd(SMDinteract.IPW.SL.b),
                        ExtractSmd(SMDinteract.IPW.SL.c))
summary(SMDinteract.IPW.SL)

+      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
+ 0.0000894 0.0387531 0.0858150 0.1054638 0.1348455 0.4105098

interact.table <- cbind(interact.table,SMDinteract.IPW.SL)
par(mar=c(2,3,2,1))
boxplot(interact.table,xlab="Approach", names=c("Unadjusted","IPW: LR", "IPW: GBM", "IPW: SL"),col='gray')
title('Balance on Interactions')
```



7. ATE Estimation

```
# Naive: simple model
coef(summary(lm(BPSysAve~SmokeNow,data=small.nhanes)))[2,]  ## -3.68
```

```
+      Estimate    Std. Error       t value      Pr(>|t|)
+ -3.6793569602  0.9639505436 -3.8169561548  0.0001411118

#Regression
coef(summary(lm(BPSysAve~SmokeNow+Gender+Age+Race3+
                         Education+MaritalStatus+HHIncome+Poverty,data=small.nhanes)))[2,]  ## -1.10

+   Estimate Std. Error    t value   Pr(>|t|)
+ -1.0977684  0.9304200 -1.1798633  0.2382629

#PS regression: quintiles
coef(summary(lm(BPSysAve~SmokeNow+ps.lr.quints,data=small.nhanes)))[2,]  ## -1.41

+   Estimate Std. Error    t value   Pr(>|t|)
+ -1.4107415  1.0450210 -1.3499648  0.1772501

#PS Regression
coef(summary(lm(BPSysAve~SmokeNow+ps.lr,data=small.nhanes)))[2,]  ## -1.11

+   Estimate Std. Error    t value   Pr(>|t|)
+ -1.1079102  1.0507005 -1.0544491  0.2918627

#PS Regression with quadratic term
coef(summary(lm(BPSysAve~SmokeNow+ps.lr+I(ps.lr^2),data=small.nhanes)))[2,]  ## -1.11

+   Estimate Std. Error    t value   Pr(>|t|)
+ -1.1103375  1.0510772 -1.0563806  0.2909802

#IPW
coef(summary(lm(BPSysAve~SmokeNow,weights=ps.lr.weight,data=small.nhanes)))[2,]  ## -1.99

+    Estimate  Std. Error     t value    Pr(>|t|)
+ -1.99123323  0.94259411 -2.11250336  0.03482317

#Matching
matched.anal <- Match(Y=small.nhanes$BPSysAve, Tr=small.nhanes$SmokeNow,
                         X=X.mat, estimand = "ATE", ties=FALSE)
matched.anal <- Match(Y=small.nhanes$BPSysAve, Tr=small.nhanes$SmokeNow,
                         X=ps.lr, estimand = "ATE", ties=FALSE)
summary(matched.anal)

+
+ Estimate...  -0.45534
+ SE.........  0.62931
+ T-stat.....  -0.72355
+ p.val......  0.46934
+
+ Original number of observations..............  1377
+ Original number of treated obs...............  595
+ Matched number of observations...............  1377
+ Matched number of observations  (unweighted).  1377

# Note that this value changes from match to match due to randomness in matching algorithm
```

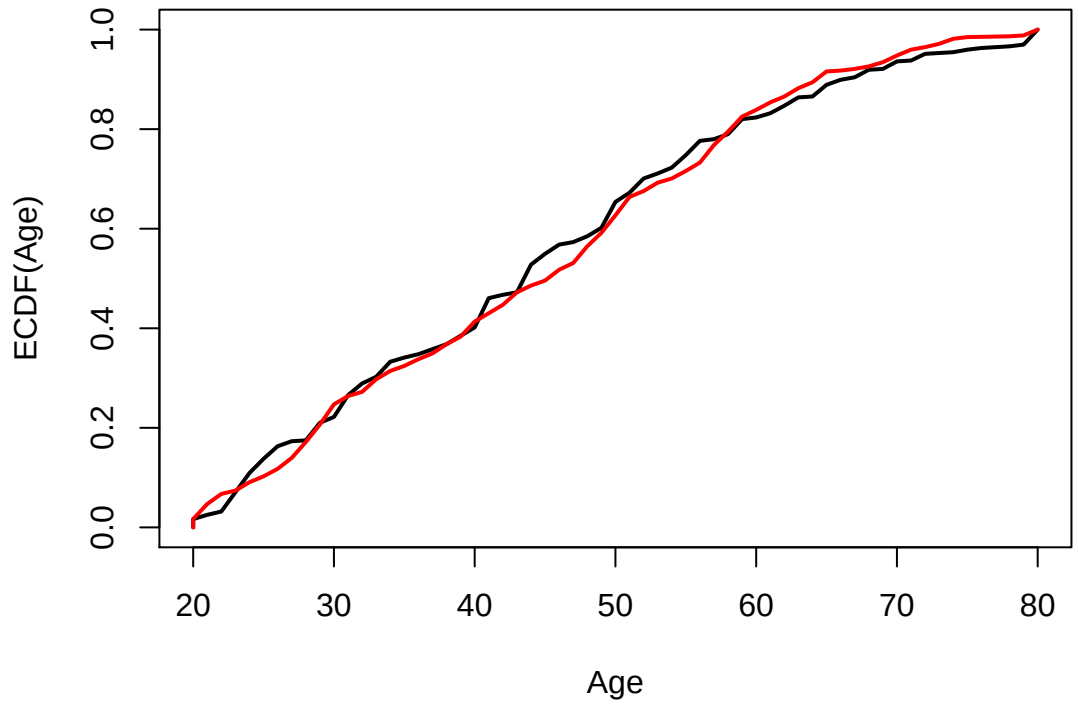
```
nhanes.allsmoke <- small.nhanes
nhanes.allsmoke$SmokeNow <- 1
nhanes.nosmoke <- small.nhanes
nhanes.nosmoke$SmokeNow <- 0
all.ests <- NULL

## Regression
```

```
coef(lm(BPSysAve~SmokeNow,data=small.nhanes))[2]

+  SmokeNow
+ -3.679357

coef(lm(BPSysAve~SmokeNow+Gender+Age+Race3+Education+MaritalStatus+HHIncome+Poverty,
                    data=small.nhanes))[2]

+  SmokeNow
+ -1.097768

## ATE via regression
mod1.lm <- lm(BPSysAve~SmokeNow+Gender+Age+Race3+Education+MaritalStatus+HHIncome+Poverty,
                    data=small.nhanes)
APO.lm.1 <- mean(predict(mod1.lm,nhanes.allsmoke))
APO.lm.0 <- mean(predict(mod1.lm,nhanes.nosmoke))
APO.lm.1 - APO.lm.0

+ [1] -1.097768

all.ests <- c(all.ests,APO.lm.1 - APO.lm.0)

mod1.lmX <- lm(BPSysAve~SmokeNow+Gender+Age+Race3+Education+MaritalStatus+HHIncome+Poverty+
                        SmokeNow:HHIncome+SmokeNow:Gender+SmokeNow:Age,data=small.nhanes)
APO.lmX.1 <- mean(predict(mod1.lmX,nhanes.allsmoke))
APO.lmX.0 <- mean(predict(mod1.lmX,nhanes.nosmoke))
APO.lmX.1 - APO.lmX.0

+ [1] -1.402538

## ATE via PS stratification
ps.lr.quints <- cut(ps.lr,quints,labels=1:5)
sum(table(cut(ps.lr,quints)))

+ [1] 1377

summary(ps.lr.quints)

+   1   2   3   4   5
+ 278 276 272 278 273

small.nhanes$ps.lr.quints <- ps.lr.quints
p.strat <- table(ps.lr.quints)/length(ps.lr.quints)
p.strat

+ ps.lr.quints
+         1         2         3         4         5
+ 0.2018882 0.2004357 0.1975309 0.2018882 0.1982571

ATE.strat <- rep(NA,5)
for(j in 1:5) {
   ATE.strat[j] <- mean(small.nhanes$BPSysAve[Smoke == 1 & small.nhanes$ps.lr.quints==j]) -
                               mean(small.nhanes$BPSysAve[Smoke == 0 & small.nhanes$ps.lr.quints==j])
}
ATE.strat

+ [1] -8.1736207 -2.2701785 -0.2062732 -1.1820287  2.8633845

sum(ATE.strat*p.strat)

+ [1] -1.816879
```

```
all.ests <- c(all.ests,sum(ATE.strat*p.strat))

## ATE via matching
mean(matched.samp$BPSysAve[matched.samp$SmokeNow == 1]) -
   mean(matched.samp$BPSysAve[matched.samp$SmokeNow == 0])

+ [1] 0.6840336

all.ests <- c(all.ests,mean(matched.samp$BPSysAve[matched.samp$SmokeNow == 1]) -
   mean(matched.samp$BPSysAve[matched.samp$SmokeNow == 0]))

## ATE via PS regression
mod1.PSlm1 <- lm(BPSysAve~SmokeNow+ps.lr,data=small.nhanes)
APO.PSlm1.1 <- mean(predict(mod1.PSlm1,nhanes.allsmoke))
APO.PSlm1.0 <- mean(predict(mod1.PSlm1,nhanes.nosmoke))
APO.PSlm1.1 - APO.PSlm1.0

+ [1] -1.10791

mod1.PSlm2 <- lm(BPSysAve~SmokeNow+ps.lr+I(ps.lr^2),data=small.nhanes)
APO.PSlm2.1 <- mean(predict(mod1.PSlm2,nhanes.allsmoke))
APO.PSlm2.0 <- mean(predict(mod1.PSlm2,nhanes.nosmoke))
APO.PSlm2.1 - APO.PSlm2.0

+ [1] -1.110337

mod1.PSlm3 <- lm(BPSysAve~SmokeNow+bs(ps.lr,df=4),data=small.nhanes)
APO.PSlm3.1 <- mean(predict(mod1.PSlm3,nhanes.allsmoke))
APO.PSlm3.0 <- mean(predict(mod1.PSlm3,nhanes.nosmoke))
APO.PSlm3.1 - APO.PSlm3.0

+ [1] -1.133493

all.ests <- c(all.ests,APO.PSlm3.1 - APO.PSlm3.0)


## ATE via IPW
small.nhanes$ps.lr.weight <- Smoke/small.nhanes$ps.lr + (1-Smoke)/(1-small.nhanes$ps.lr)
IPW.est<-mean(Smoke*small.nhanes$BPSysAve*small.nhanes$ps.lr.weight) -
            mean((1-Smoke)*small.nhanes$BPSysAve*small.nhanes$ps.lr.weight)
all.ests <- c(all.ests,IPW.est)


coef(lm(BPSysAve ~ SmokeNow, weights = ps.lr.weight,data=small.nhanes))

+ (Intercept)    SmokeNow
+   124.237219   -1.991233

mean(Smoke*small.nhanes$BPSysAve/small.nhanes$ps.lr) -
    mean((1-Smoke)*small.nhanes$BPSysAve/(1-small.nhanes$ps.lr))

+ [1] -1.928655

# Final table:
round(rbind(Max.SMD, Mean.SMD, Med.SMD, all.ests), 3)

+            [,1]   [,2]  [,3]   [,4]   [,5]
+ Max.SMD      NA  0.592 0.185     NA  0.052
+ Mean.SMD     NA  0.254 0.111     NA  0.024
+ Med.SMD      NA  0.248 0.115     NA  0.023
+ all.ests -1.098 -1.817 0.684 -1.133 -1.929
```

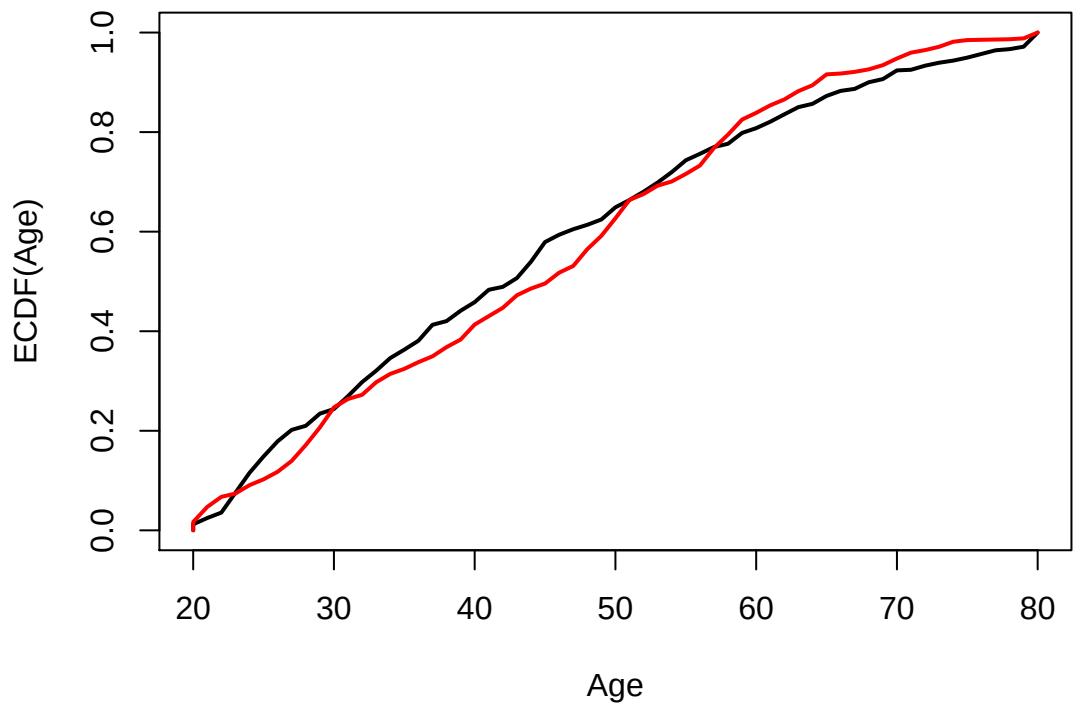
8. ATT Estimation

```
matched.anal.ATT <- Match(Y=small.nhanes$BPSysAve, Tr=small.nhanes$SmokeNow, X=ps.lr,
                          estimand = "ATT", ties=FALSE)
summary(matched.anal.ATT)

+
+ Estimate...  0.75462
+ SE.........  0.94712
+ T-stat.....  0.79676
+ p.val......  0.42559
+
+ Original number of observations..............  1377
+ Original number of treated obs...............  595
+ Matched number of observations...............  595
+ Matched number of observations  (unweighted).  595

matched.samp.ATT <- small.nhanes[c(matched.anal.ATT$index.control,matched.anal.ATT$index.treated),]
mean(matched.samp.ATT$BPSysAve[matched.samp.ATT$SmokeNow == 1]) -
   mean(matched.samp.ATT$BPSysAve[matched.samp.ATT$SmokeNow == 0])

+ [1] 0.7546218

temp0 <- Ecdf(matched.samp.ATT$Age[matched.samp.ATT$SmokeNow==0],pl=F)
temp1 <- Ecdf(matched.samp.ATT$Age[matched.samp.ATT$SmokeNow==1],pl=F)

par(mar=c(4,4,2,1))

plot(temp0$x,temp0$y,ylab="ECDF(Age)",xlab="Age",main="",type="l",lwd=2)
lines(temp1$x,temp1$y,col="red",lwd=2)
```




```
ATT.match <- CreateTableOne(vars = vars, strata = "SmokeNow", data = matched.samp.ATT, test = FALSE)
SMD.ATT <- ExtractSmd(ATT.match)
```

```r
## ATT via IPW
small.nhanes$ATT.lr.weight <- Smoke + (1-Smoke)*ps.lr/(1-ps.lr)
temp0 <- Ecdf(small.nhanes$Age[Smoke==0],weights=small.nhanes$ATT.lr.weight[Smoke==0],pl=F)
temp1 <- Ecdf(small.nhanes$Age[Smoke==1],weights=small.nhanes$ATT.lr.weight[Smoke==1],pl=F)
plot(temp0$x,temp0$y,ylab="ECDF(Age)",xlab="Age",main="",type="l",lwd=2)
lines(temp1$x,temp1$y,col="red",lwd=2)
```




```r
nhanes.ATT.IPW <- svydesign(ids=~0, data=small.nhanes, weights=small.nhanes$ATT.lr.weight)
ATT.IPW <- svyCreateTableOne(vars = vars, strata = "SmokeNow", data = nhanes.ATT.IPW, test = FALSE)
print(ATT.IPW, smd = TRUE)
```

```
+                     Stratified by SmokeNow
+                      0              1              SMD
+   n                   597.0          595.0
+   Gender = male (%)   365.4 (61.2)   369.0 (62.0)   0.016
+   Age (mean (sd))     44.46 (16.35)  44.96 (15.11)  0.031
+   Race3 (%)                                         0.110
+      Asian             15.4 ( 2.6)    15.0 ( 2.5)
+      Black             76.0 (12.7)    64.0 (10.8)
+      Hispanic          49.5 ( 8.3)    38.0 ( 6.4)
+      Mexican           36.2 ( 6.1)    35.0 ( 5.9)
+      White            389.8 (65.3)   416.0 (69.9)
+      Other             30.0 ( 5.0)    27.0 ( 4.5)
+   Education (%)                                     0.065
+      8th Grade         30.4 ( 5.1)    33.0 ( 5.5)
+      9 - 11th Grade   113.6 (19.0)   120.0 (20.2)
+      High School      141.9 (23.8)   151.0 (25.4)
+      Some College     226.7 (38.0)   210.0 (35.3)
+      College Grad      84.4 (14.1)    81.0 (13.6)
+   MaritalStatus (%)                                 0.034
+      Divorced          76.0 (12.7)    77.0 (12.9)
+      LivePartner       93.6 (15.7)    96.0 (16.1)
+      Married          242.0 (40.5)   240.0 (40.3)
```

```
+      NeverMarried      141.9 (23.8)   142.0 (23.9)
+      Separated          17.1 ( 2.9)    14.0 ( 2.4)
+      Widowed            26.4 ( 4.4)    26.0 ( 4.4)
+   Poverty (mean (sd))   2.39 (1.60)    2.38 (1.58)   0.006

round(cbind(SMD.ATT,ExtractSmd(ATT.IPW)),3)

+               SMD.ATT
+ Gender          0.123 0.016
+ Age             0.009 0.031
+ Race3           0.174 0.110
+ Education       0.090 0.065
+ MaritalStatus   0.171 0.034
+ Poverty         0.075 0.006

mean(Smoke*small.nhanes$BPSysAve*small.nhanes$ATT.lr.weight) -
   mean((1-Smoke)*small.nhanes$BPSysAve*small.nhanes$ATT.lr.weight)

+ [1] -0.3895692
```