Causal_project

2022-11-15

Install packages & library

```
#install.packages("remotes")
#remotes::install_github("ygeunkim/propensityml")
#install.packages("dplyr")
#install.packages("sas7bdat")

library(propensityml)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(sas7bdat)
```

Korea National Health and Nutrition Examination Survey Dataset

1. Data reading

```
raw_data <- read.sas7bdat("/Users/kyungseonlee/snu-causal/main/input/rawdata/HN20_AL
L.sas7bdat")
```

```
## Warning in gsub("^ +| +$", "", data[[col$name]][row]): unpaired surrogate
## Unicode point de37

## Warning in gsub("^ +| +$", "", data[[col$name]][row]): unpaired surrogate
## Unicode point de37
## Warning in gsub("^ +| +$", "", data[[col$name]][row]): unpaired surrogate
## Unicode point dd75

## Warning in gsub("^ +| +$", "", data[[col$name]][row]): unpaired surrogate
## Unicode point dd75
```

```
## Warning in gsub("^ +| +$", "", data[[col$name]][row]): unpaired surrogate
## Unicode point de37

## Warning in gsub("^ +| +$", "", data[[col$name]][row]): unpaired surrogate
## Unicode point de37
```

```
## Warning in gsub("^ +| +$", "", data[[col$name]][row]): unpaired surrogate
## Unicode point dff1

## Warning in gsub("^ +| +$", "", data[[col$name]][row]): unpaired surrogate
## Unicode point dff1
```

```
## Warning in gsub("^ +| +$", "", data[[col$name]][row]): unpaired surrogate
## Unicode point de37

## Warning in gsub("^ +| +$", "", data[[col$name]][row]): unpaired surrogate
## Unicode point de37

## Warning in gsub("^ +| +$", "", data[[col$name]][row]): unpaired surrogate
## Unicode point de37

## Warning in gsub("^ +| +$", "", data[[col$name]][row]): unpaired surrogate
## Unicode point de37

## Warning in gsub("^ +| +$", "", data[[col$name]][row]): unpaired surrogate
## Unicode point de37

## Warning in gsub("^ +| +$", "", data[[col$name]][row]): unpaired surrogate
## Unicode point de37
```

```
## Warning in gsub("^ +| +$", "", data[[col$name]][row]): unpaired surrogate
## Unicode point dda7

## Warning in gsub("^ +| +$", "", data[[col$name]][row]): unpaired surrogate
## Unicode point dda7
```

```
## Warning in gsub("^ +| +$", "", data[[col$name]][row]): unpaired surrogate
## Unicode point dc78

## Warning in gsub("^ +| +$", "", data[[col$name]][row]): unpaired surrogate
## Unicode point dc78
```

```
dim(raw_data) %>% head()
```

```
## [1] 7359  762
```

```
#raw_data %>% head(1)
```

2. Data preprocessing

```
column=c('sex','age','BD1','HE_ast','HE_alt',"HE_BMI",'DK8_dg','DK9_dg')
#column2=c('BD1','DC2_dg','DK8_dg','DK9_dg','DK4_dg')

df=raw_data[,column]
df %>% head(7)
```

```
##   sex age BD1 HE_ast HE_alt   HE_BMI DK8_dg DK9_dg
## 1   1  39   2     26     23 24.18549      0      0
## 2   2  39   2     22     20 17.93594      0      0
## 3   1  10   8     32     15 13.99727      8      8
## 4   1   7   8    NaN    NaN 16.51734      8      8
## 5   1   4   8    NaN    NaN 14.09464      8      8
## 6   1  60   1     30     25      NaN      9      9
## 7   2  58   2     28     33 26.58997      0      0
```

```
sum(df[,c(8)])
```

```
## [1] NaN
```

Missing value deletion

```
for(i in 3:ncol(df)){
  if(i==4|i==5|i==6){
    df<- df[!( is.na(df[,i])), ]
  }else{
    df<- df[!(df[,i] == 8 |df[,i] == 9 | is.na(df[,i])), ]
  }
}
for (i in 1:nrow(df)){
  if (df[i,7]+df[i,8]==0){df[i,9]=0}else{df[i,9]=1}
}
df %>% head(20)
```

```
##    sex age BD1 HE_ast HE_alt   HE_BMI DK8_dg DK9_dg V9
## 1    1  39   2     26     23 24.18549      0      0  0
## 2    2  39   2     22     20 17.93594      0      0  0
## 7    2  58   2     28     33 26.58997      0      0  0
## 8    1  56   2     28     25 23.68213      0      0  0
## 9    2  53   2     25     16 19.66942      0      0  0
## 10   1  20   2     23     27 20.84331      0      0  0
## 11   1  24   2     20     24 21.04169      0      0  0
## 12   1  56   2     27     31 26.96156      0      0  0
## 13   2  53   2     24     19 24.03803      0      0  0
## 16   2  74   2     26     18 27.06330      0      0  0
## 17   1  51   2     32     45 23.85387      0      0  0
## 18   2  47   2     26     25 26.91985      0      0  0
## 19   2  19   2     31     13 15.99924      0      0  0
## 21   1  67   2     24     22 24.55823      0      0  0
## 22   2  65   2     23     24 22.76996      0      0  0
## 23   2  39   2     14      8 18.94065      0      0  0
## 24   1  41   2     21     21 24.81843      0      0  0
## 26   1  60   2     34     20 25.08102      0      0  0
## 27   2  56   2     26     40 23.35884      0      0  0
## 28   1  28   2     20     15 27.68546      0      0  0
```

```
sum(is.na(df))
```

```
## [1] 0
```

```
dim(df)
```

```
## [1] 5265    9
```

```
# First topic: Hepatitis causal inference
df1=df[,c(-7,-8)]
colnames(df1)[3:ncol(df1)]=c("treat","ast","alt","bmi","outcome")

df1 %>% head(7)
```

```
##    sex age treat ast alt      bmi outcome
## 1    1  39     2  26  23 24.18549       0
## 2    2  39     2  22  20 17.93594       0
## 7    2  58     2  28  33 26.58997       0
## 8    1  56     2  28  25 23.68213       0
## 9    2  53     2  25  16 19.66942       0
## 10   1  20     2  23  27 20.84331       0
## 11   1  24     2  20  24 21.04169       0
```

Diabetes & Drinking

```
df1$treat<-df1$treat-1
df1$sex<-df1$sex-1

df1 %>% head(7)
```

```
##    sex age treat ast alt      bmi outcome
## 1    0  39     1  26  23 24.18549       0
## 2    1  39     1  22  20 17.93594       0
## 7    1  58     1  28  33 26.58997       0
## 8    0  56     1  28  25 23.68213       0
## 9    1  53     1  25  16 19.66942       0
## 10   0  20     1  23  27 20.84331       0
## 11   0  24     1  20  24 21.04169       0
```

```
table(df1$treat,df1$outcome)
```

```
##
##         0    1
##   0   558    7
##   1  4642   58
```

```
# (df1$DC8_dg==0)
# sum(is.na(df1$treat))
# sum(is.na(df1$outcome))
# dim(df1)
# which(df1$treat==0) %>% head()
```

Propensity score estimation- logistic regression, random Forest, CART

1. Logistic Regression

1-a. LR-ps estimation

```
#logsitic regression
df1_lo=df1
df1_lo %>% head(10)
```

```
##    sex age treat ast alt      bmi outcome
## 1    0  39     1  26  23 24.18549       0
## 2    1  39     1  22  20 17.93594       0
## 7    1  58     1  28  33 26.58997       0
## 8    0  56     1  28  25 23.68213       0
## 9    1  53     1  25  16 19.66942       0
## 10   0  20     1  23  27 20.84331       0
## 11   0  24     1  20  24 21.04169       0
## 12   0  56     1  27  31 26.96156       0
## 13   1  53     1  24  19 24.03803       0
## 16   1  74     1  26  18 27.06330       0
```

```
log_reg_ps=glm(treat~ .-outcome, family = "binomial", data=df1)
logit_e_hat=predict(log_reg_ps)
print(logit_e_hat %>% head())
```

```
##        1        2        7        8        9       10 
## 4.527971 2.821645 1.460026 3.419172 1.983010 5.704291
```

```
lo_ps=exp(logit_e_hat)/(1+exp(logit_e_hat))
print(lo_ps %>% head())
```

```
##         1         2         7         8         9        10 
## 0.9893129 0.9438343 0.8115367 0.9682984 0.8790017 0.9966794
```

```
df1_lo[,"ps"]=lo_ps
#
# for (i in 1:nrow(df1_lo) ){
#   if (df1_lo$treat[i]==1){
#     df1_lo[i,"ps"]=lo_ps[i]
#   }
#   else{
#     df1_lo[i,"ps"]=lo_ps[i]
#   }
# }
df1_lo %>% head()
```

```
##    sex age treat ast alt      bmi outcome        ps
## 1    0  39     1  26  23 24.18549       0 0.9893129
## 2    1  39     1  22  20 17.93594       0 0.9438343
## 7    1  58     1  28  33 26.58997       0 0.8115367
## 8    0  56     1  28  25 23.68213       0 0.9682984
## 9    1  53     1  25  16 19.66942       0 0.8790017
## 10   0  20     1  23  27 20.84331       0 0.9966794
```

1-b. LR-weighting

```
zi=df1_lo$treat
yi=df1_lo$outcome
e=df1_lo$ps

df1_lo["ipw_wt"]= zi/e-(1-zi)/(1-e)
df1_lo %>% head()
```

```
##    sex age treat ast alt      bmi outcome        ps   ipw_wt
## 1    0  39     1  26  23 24.18549       0 0.9893129 1.010803
## 2    1  39     1  22  20 17.93594       0 0.9438343 1.059508
## 7    1  58     1  28  33 26.58997       0 0.8115367 1.232230
## 8    0  56     1  28  25 23.68213       0 0.9682984 1.032740
## 9    1  53     1  25  16 19.66942       0 0.8790017 1.137654
## 10   0  20     1  23  27 20.84331       0 0.9966794 1.003332
```

```
ATE_ipw_log=mean(zi*yi/e)-mean((1-zi)*yi/(1-e))
ATE_ipw_log
```

```
## [1] 0.006323189
```

```
ATE_sipw_log=sum(zi*yi/e)/sum(zi/e)-sum((1-zi)*yi/(1-e))/sum((1-zi)/(1-e))
ATE_sipw_log
```

```
## [1] 0.007927617
```

1-c. LR-Evaluate

```
cov_balance_lo=data.frame(rep(0),row.names = "logistic regression")

for(i in colnames(df1)){
  if(i!="treat" & i!="outcome"){
#    print(df1[i])
    t_weighted_mean=mean((df1[i]*df1_lo$ipw_wt)[df1$treat==1,])
    c_weighted_mean=mean((df1[i]*df1_lo$ipw_wt)[df1$treat==0,])
    weighted_mean_diff=abs(t_weighted_mean-c_weighted_mean)
    asam=weighted_mean_diff/sd((df1[i]*df1_lo$ipw_wt)[df1$treat==1,])
    cov_balance_lo[i]=asam
  }
}
cov_balance_lo=cov_balance_lo[,-1]
cov_balance_lo$ASAM=apply(cov_balance_lo,1,mean)
cov_balance_lo["ASAM"]
```

```
##                          ASAM
## logistic regression 24.38727
```

2. Random Forest

2-a. RF-ps estimation

```
library(randomForest)
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
df1_rf=df1

rf=randomForest(treat~ .-outcome, data=df1)
```

```
## Warning in randomForest.default(m, y, ...): The response has five or fewer
## unique values. Are you sure you want to do regression?
```

```
rf_ps=predict(rf)
print(rf_ps %>% head())
```

```
##         1         2         7         8         9        10
## 0.9819949 0.9291145 0.8726215 0.9835696 0.8906738 0.9650982
```

```
df1_rf[,"ps"]=rf_ps

# for (i in 1:nrow(df1_rf) ){
#   if (df1_rf$treat[i]==1){
#     df1_rf[i,"ps"]=rf_ps[i]
#   }
#   else{
#     df1_rf[i,"ps"]=rf_ps[i]
#   }
# }
df1_rf %>% head()
```

```
##    sex age treat ast alt      bmi outcome        ps
## 1    0  39     1  26  23 24.18549       0 0.9819949
## 2    1  39     1  22  20 17.93594       0 0.9291145
## 7    1  58     1  28  33 26.58997       0 0.8726215
## 8    0  56     1  28  25 23.68213       0 0.9835696
## 9    1  53     1  25  16 19.66942       0 0.8906738
## 10   0  20     1  23  27 20.84331       0 0.9650982
```

2-b. RF-weighting

```
zi=df1_rf$treat
yi=df1_rf$outcome
e=df1_rf$ps

df1_rf["ipw_wt"]= zi/e-(1-zi)/(1-e)
df1_rf %>% head()
```

```
##    sex age treat ast alt      bmi outcome        ps   ipw_wt
## 1    0  39     1  26  23 24.18549       0 0.9819949 1.018335
## 2    1  39     1  22  20 17.93594       0 0.9291145 1.076294
## 7    1  58     1  28  33 26.58997       0 0.8726215 1.145972
## 8    0  56     1  28  25 23.68213       0 0.9835696 1.016705
## 9    1  53     1  25  16 19.66942       0 0.8906738 1.122745
## 10   0  20     1  23  27 20.84331       0 0.9650982 1.036164
```

```
#ATE estimation
ATE_ipw_rf=mean(zi*yi/e)-mean((1-zi)*yi/(1-e))
ATE_ipw_rf
```

```
## [1] 0.005680105
```

```
ATE_sipw_rf=sum(zi*yi/e)/sum(zi/e)-sum((1-zi)*yi/(1-e))/sum((1-zi)/(1-e))
ATE_sipw_rf
```

```
## [1] 0.005110936
```

2-c. RF-Evaluation

```
cov_balance_rf=data.frame(rep(0),row.names = "Random Forest")

for(i in colnames(df1)){
  if(i!="treat" & i!="outcome"){
#    print(df1[i])
    t_weighted_mean=mean((df1[i]*df1_rf$ipw_wt)[df1$treat==1,])
    c_weighted_mean=mean((df1[i]*df1_rf$ipw_wt)[df1$treat==0,])
    weighted_mean_diff=abs(t_weighted_mean-c_weighted_mean)
    asam=weighted_mean_diff/sd((df1[i]*df1_rf$ipw_wt)[df1$treat==1,])
    cov_balance_rf[i]=asam
  }
}
cov_balance_rf=cov_balance_rf[,-1]
cov_balance_rf
```

```
##                      sex      age      ast      alt      bmi
## Random Forest 8.856167 21.43868 12.27349 10.59225 44.42541
```

```
cov_balance_rf$ASAM=apply(cov_balance_rf,1,mean)
cov_balance_rf["ASAM"]
```

```
##                    ASAM
## Random Forest 19.5172
```

3. CART

3-a. CART-ps estimation

```
#install.packages("rpart")
library(rpart)

# cart=rpart(poisox~ .-outcome, data=df10 ,method='poisson')
# summary(cart)
#
# df10_c=df10
# cart_ps=predict(cart)
# df10_c["propensity score"]=cart_ps
# df10_c %>% head()
#
# df11_c=df11

df1_c=df1
df1_c %>% head(5)
```

```
##   sex age treat ast alt      bmi outcome
## 1   0  39     1  26  23 24.18549       0
## 2   1  39     1  22  20 17.93594       0
## 7   1  58     1  28  33 26.58997       0
## 8   0  56     1  28  25 23.68213       0
## 9   1  53     1  25  16 19.66942       0
```

```
cart=rpart(treat~ .-outcome, data=df1_c ,method='poisson',control = rpart.control(max
depth = 5))
#summary(cart2)
cart$cptable
```

```
##           CP nsplit rel error    xerror       xstd
## 1 0.04193620      0 1.0000000 1.0000981 0.03731780
## 2 0.03490859      1 0.9580638 0.9584365 0.03454442
## 3 0.01000000      2 0.9231552 0.9236835 0.03287868
```

```
cart_ps=predict(cart)
cart_ps %>% head()
```

```
##         1         2         7         8         9        10
## 0.9510929 0.9510929 0.9510929 0.9510929 0.9510929 0.9510929
```

```
df1_c[,"ps"]=cart_ps

df1_c %>% head()
```

```
##    sex age treat ast alt      bmi outcome        ps
## 1    0  39     1  26  23 24.18549       0 0.9510929
## 2    1  39     1  22  20 17.93594       0 0.9510929
## 7    1  58     1  28  33 26.58997       0 0.9510929
## 8    0  56     1  28  25 23.68213       0 0.9510929
## 9    1  53     1  25  16 19.66942       0 0.9510929
## 10   0  20     1  23  27 20.84331       0 0.9510929
```

```
sum(df1_c$ps==0|df1_c$ps==1)
```

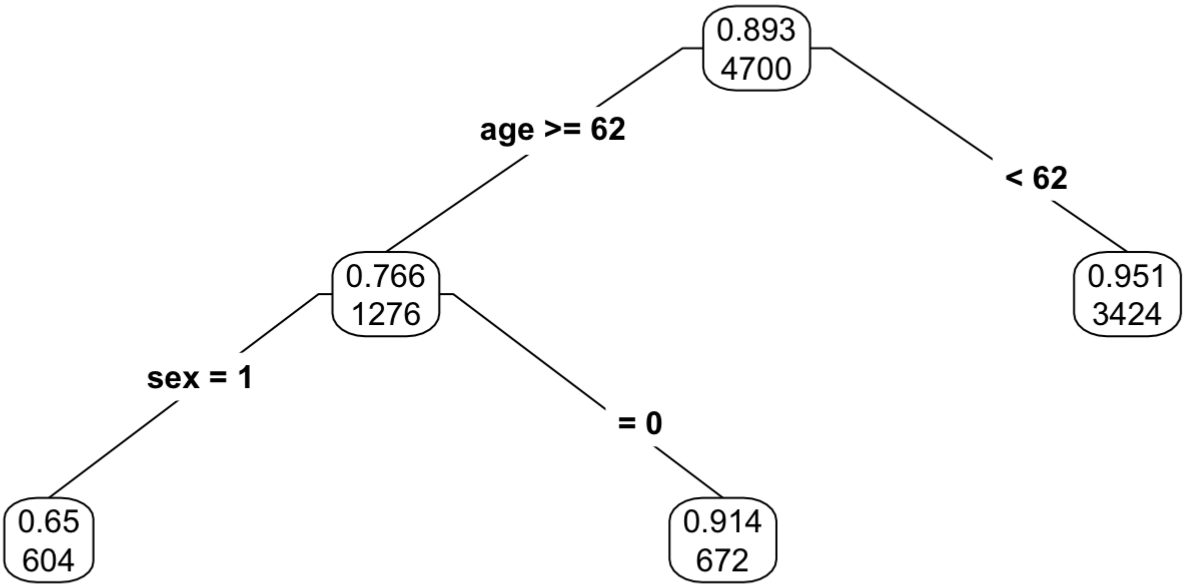
```
## [1] 0
```

```
#df1_c[65:70,]
```

```
#install.packages("rpart.plot")

library(rpart.plot)

prp(cart, type=4, extra=2, digits=3)
```



### 3-b. CART-weighting

```
#add weighting- CART
zi=df1_c$treat
yi=df1_c$outcome
e=df1_c$ps

df1_c["ipw_wt"]= zi/e-(1-zi)/(1-e)
df1_c %>% head()
```

```
##    sex age treat ast alt      bmi outcome        ps   ipw_wt
## 1    0  39     1  26  23 24.18549       0 0.9510929 1.051422
## 2    1  39     1  22  20 17.93594       0 0.9510929 1.051422
## 7    1  58     1  28  33 26.58997       0 0.9510929 1.051422
## 8    0  56     1  28  25 23.68213       0 0.9510929 1.051422
## 9    1  53     1  25  16 19.66942       0 0.9510929 1.051422
## 10   0  20     1  23  27 20.84331       0 0.9510929 1.051422
```

```
#ATE_ipw
ATE_ipw_cart=mean(zi*yi/e)-mean((1-zi)*yi/(1-e))
ATE_ipw_cart
```

```
## [1] 0.001957638
```

```
#ATE_sipw
ATE_sipw_cart=sum(zi*yi/e)/sum(zi/e)-sum((1-zi)*yi/(1-e))/sum((1-zi)/(1-e))
ATE_sipw_cart
```

```
## [1] 0.001956726
```

3-c. CART-evaluation

```
cov_balance_cart=data.frame(rep(0),row.names = "CART")

for(i in colnames(df1)){
  if(i!="treat" & i!="outcome"){
#    print(df1[i])
    t_weighted_mean=mean((df1[i]*df1_c$ipw_wt)[df1$treat==1,])
    c_weighted_mean=mean((df1[i]*df1_c$ipw_wt)[df1$treat==0,])
    weighted_mean_diff=abs(t_weighted_mean-c_weighted_mean)
    asam=weighted_mean_diff/sd((df1[i]*df1_c$ipw_wt)[df1$treat==1,])
    cov_balance_cart[i]=asam
  }
}
cov_balance_cart=cov_balance_cart[,-1]
cov_balance_cart$ASAM=apply(cov_balance_cart,1,mean)
cov_balance_cart["ASAM"]
```

```
##          ASAM
## CART 20.66538
```

Total ATE table

```
ATE_table= rbind(ATE_ipw_log,
      ATE_sipw_log,
      ATE_ipw_rf,
      ATE_sipw_rf,
      ATE_ipw_cart,
      ATE_sipw_cart)

colnames(ATE_table)="ATE table in KHN Dataset"

knitr :: kable(ATE_table,"simple")
```

	ATE table in KHN Dataset
ATE_ipw_log	0.0063232
ATE_sipw_log	0.0079276
ATE_ipw_rf	0.0056801
ATE_sipw_rf	0.0051109
ATE_ipw_cart	0.0019576
ATE_sipw_cart	0.0019567

```
#install.packages("knitr")
```

Evaluation visualization - Chemical dataset

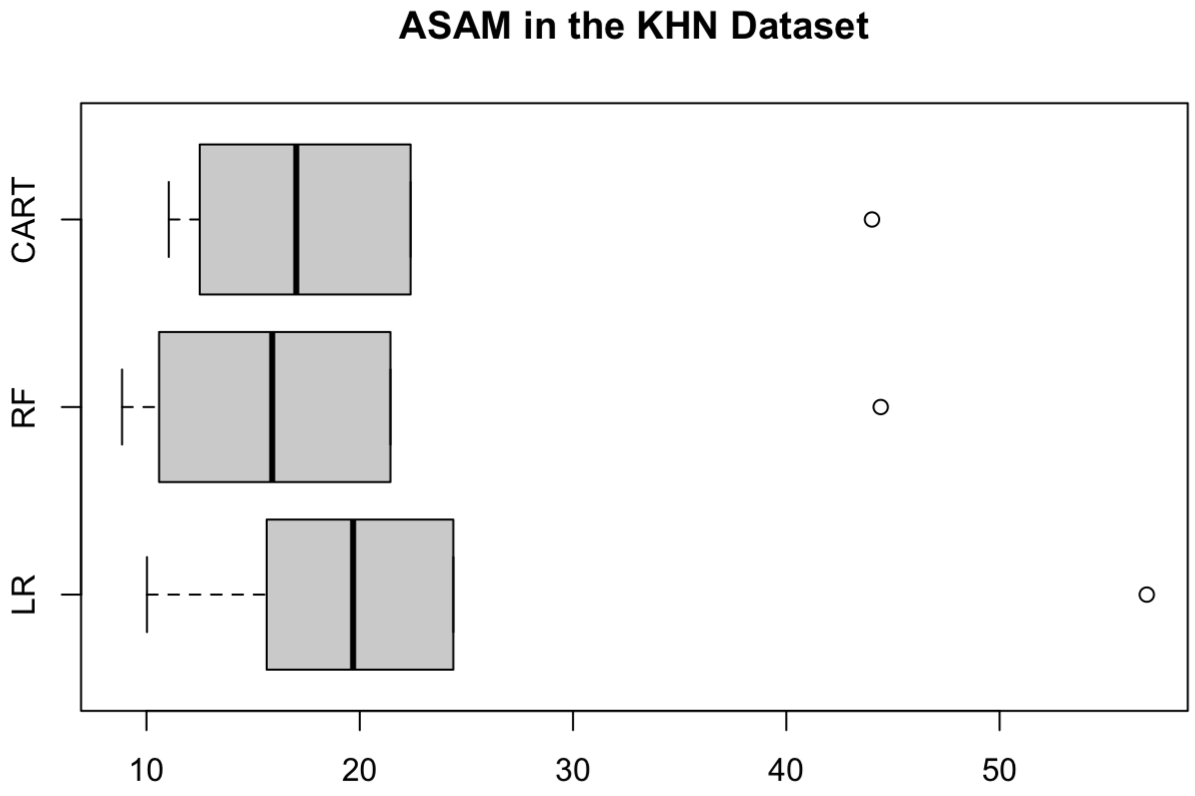
a. ASAM table

```
ASAM_table2=rbind(cov_balance_lo["ASAM"] ,cov_balance_rf["ASAM"] ,cov_balance_cart["A
SAM"])
colnames(ASAM_table2)="ASAM in KHN Dataset"
knitr::kable(ASAM_table2,"simple")
```

	ASAM in KHN Dataset
logistic regression	24.38727
Random Forest	19.51720
CART	20.66538

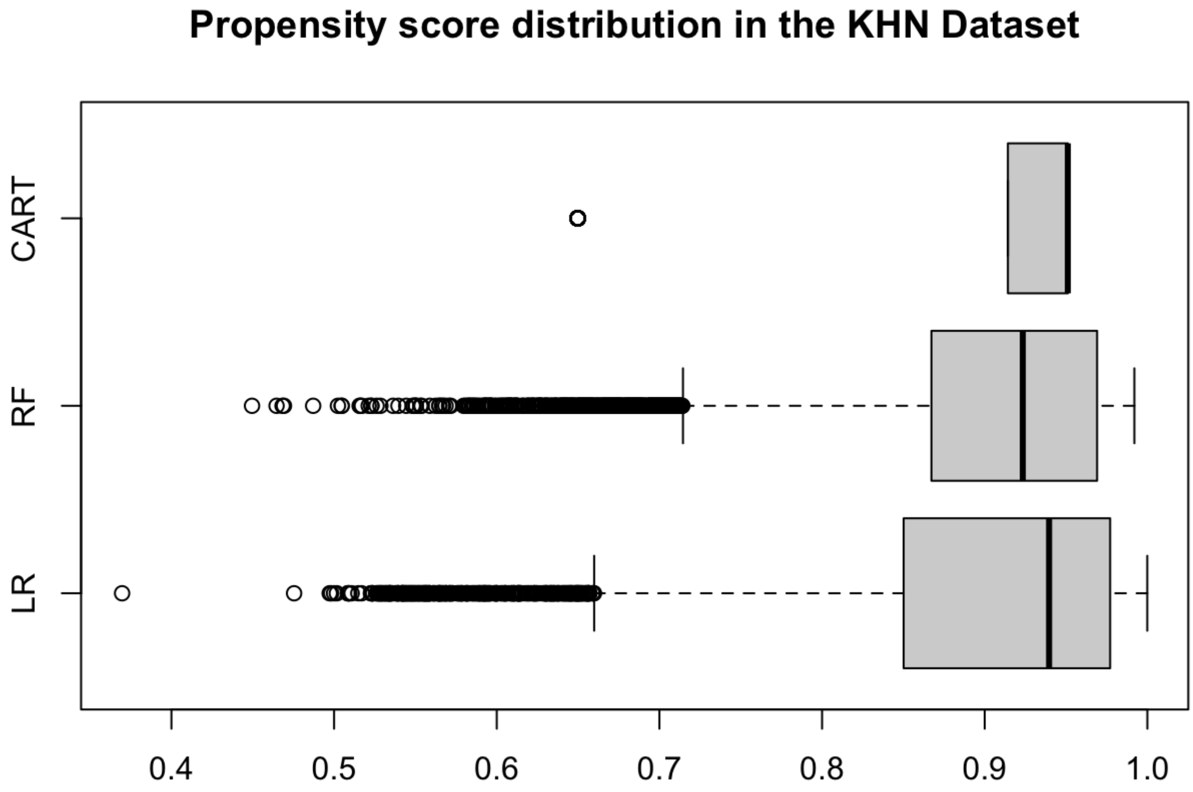
b. ASAM box plot

```
a=cbind(t(cov_balance_lo) ,t(cov_balance_rf) ,t(cov_balance_cart))
colnames(a)=c("LR","RF","CART")
boxplot(a, main="ASAM in the KHN Dataset",horizontal = TRUE)
```



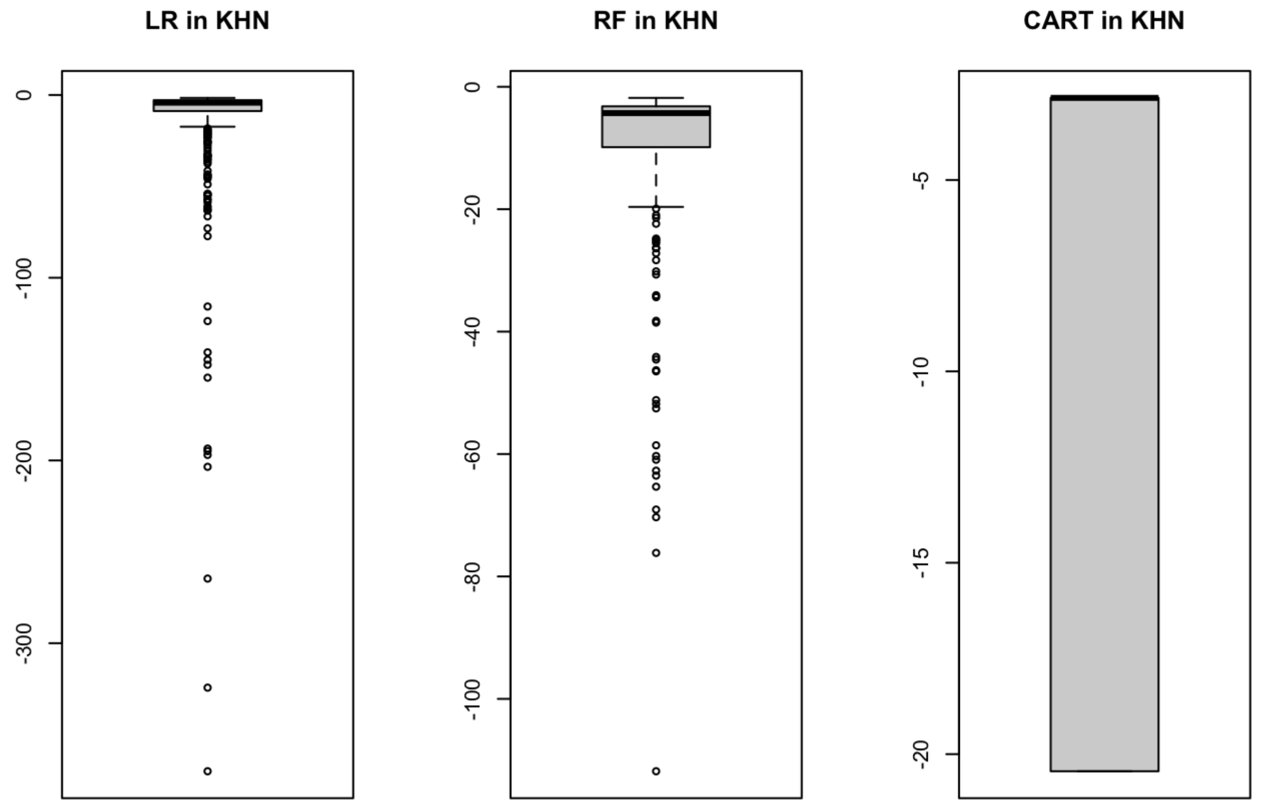
c. ps distribution

```
b=cbind(df1_lo["ps"],df1_rf["ps"],df1_c["ps"])
colnames(b)=c("LR","RF","CART")
boxplot(b,horizontal = TRUE,main="Propensity score distribution in the KHN Dataset")
```



d. weight distribution

```
par(mfcol=c(1,3))
boxplot(df1_lo[(df1_lo$treat==0),"ipw_wt"],main="LR in KHN")
boxplot(df1_rf[(df1_lo$treat==0),"ipw_wt"],main="RF in KHN")
boxplot(df1_c[(df1_lo$treat==0),"ipw_wt"],main="CART in KHN")
```



```
sim=data.frame(c(0.094,0.075,0.143))
colnames(sim)="ASAM in simulated dataset"
rownames(sim)=c("logistic regression","Random Forest","CART")
knitr :: kable(sim,"simple")
```

	ASAM in simulated dataset
logistic regression	0.094
Random Forest	0.075
CART	0.143