Lab: Propensity score matching and weighting (binary treatment)

Fan Li

Contents

0.1 Setup: Loading Packages

```
library(PSweight)
library(Matching)
library(glmnet)
library(progress)
```

1 Binary treatment example: RHC data

This lab shows several methods for estimating causal effects from observational studies: outcome modeling, propensity scores, matching, weighting. For matching, we use the R package Matching; for weighting, we use the R package PSweight. We illustrate via the publicly available RHC data set.

1.1 The RHC Data Set

Right heart catheterization (RHC) is a diagnostic procedure for directly measuring cardiac function in critically ill patients. In an influential study, Connors et al. (1996) studied the effectiveness of RHC with an observational study design. The study collected data on 5735 hospitalized adult patients; 2184 of them are assigned to the treatment ($Z = 1$), receipt of RHC within 24 hours of admission, and the remaining 3551 assigned to the control condition ($Z = 0$). The outcome was survival at 30 days after admission. The goal is to assess the causal effect of RHC on the binary outcome, death at 30 days after admission.

For this lab, we have prepared a cleaned data set based on the original RHC data, availablehere. The treatment variable is swang1, the outcome variable is death. To simplify the illustration, we only restrict to 20 covariates that we identified to be top confounders (in an ad-hoc exploratory analysis – this is only to make the code run faster and by no means a recommended practice). We generically denote the confounders by X, and the observed outcome $Y = ZY(1) + (1 - Z)Y(0)$. The outcome in RHC is binary, therefore we are interested in estimating the marginal (marginal means over a specific target population) causal risk difference

(RD), risk ratio (RR), and odds ratio (OR). For each effect, we report the point estimates, the standard errors, and the corresponding 95% confidence interval.

For each causal estimand, we look at different target population: for ATE, the target population is the population represented by the sample; for ATT, the target population is the treated population; for ATO, the target population is the population where the treatment and control group are most similar (population in clinical equipoise).

```
#Load the RHC data set
rhc <- read.csv("rhc_demo.csv")[,-c(1)]
#Transfer the outcome and the treatment to binary variables
rhc$swang1 <- factor(rhc$swang1)
```

Below table summarizes the interpretations of most of the covariates (Hirno and Imbens, 2001).

Name	Interpretations
age	Age (years)
sex	Female
cat1_copd	COPD
cat1_mosfsep	MOSF w/Sepsis
cat1_mosfmal	MOSF w/Malignancy
cat1_chf	CHF
cat1_coma	Coma
cat1_cirr	Cirrhosis
cat1_lung	Lung Cancer
cat1_colon	Colon Cancer
cat2_mosfsep	MOSF w/Sepsis
cat2_coma	Coma
cat2_mosfmal	MOSF w/Malignancy
cat2_lung	Lung Cancer
cat2_cirr	Cirrhosis
cat2_colon	Colon Cancer
ca_yes	Cancer-localized
ca_meta	Cancer-metastatic
pafi1	PaO2/F102 ratio
wtkilo1	Weight
surv2md1	Estimate of prob. of surviving 2 months
dementhx	Dementia, stroke or cerebral infarct, Parkinson's disease
gastr	Gastrointestinal diagnosis
wblc1	WBC
temp1	Temperature
das2d3pc	DASI-Duke Activity Status Index
chfhx	Congestive Heart Failure
hema	Hematological diagnosis
chrpulhx	Chronic pulmonary disease, severe pulmonary disease
cardiohx	Cardiovascular symptoms
meta	Metabolic diagnosis

```
#Present the summary statistics of the outcome and the treatment
summary(rhc[,c("death","swang1")])
```

```
     death          swang1
 Min.   :0.000   No RHC:3551
 1st Qu.:0.000   RHC   :2184
```

```
 Median :1.000
 Mean   :0.649
 3rd Qu.:1.000
 Max.   :1.000
```

```
#Ensure the type of categorical variables is correct
rhc_factor <- c("cat1","cat2","ca","dementhx","gastr","chfhx","sex","hema",
                "chrpulhx","cardiohx","meta","adld3p")
rhc_rep <- rhc
rhc_rep[,rhc_factor] <- lapply(rhc_rep[,rhc_factor], factor)
#Present the summary statistics of all covariates
summary(rhc_rep[,-c(4,16)])
```

```
      ptid          survtime         cat1               ca
 Min.   :    5   Min.   :   2.0   ARF  :2490   Metastatic: 384
 1st Qu.: 2562   1st Qu.:  16.0   CHF  : 456   No         :4379
 Median : 5131   Median : 166.0   Coma : 436   Yes        : 972
 Mean   : 5134   Mean   : 186.4   COPD : 457
 3rd Qu.: 7689   3rd Qu.: 232.0   MOSF :1626
 Max.   :10278   Max.   :1943.0   Other: 270

     death       cardiohx chfhx    dementhx chrpulhx      age
 Min.   :0.000   0:4722   0:4714   0:5171   0:4646   Min.   : 18.04
 1st Qu.:0.000   1:1013   1:1021   1: 564   1:1089   1st Qu.: 50.15
 Median :1.000                                      Median : 64.05
 Mean   :0.649                                      Mean   : 61.38
 3rd Qu.:1.000                                      3rd Qu.: 73.93
 Max.   :1.000                                      Max.   :101.85

     sex          surv2md1         das2d3pc         wblc1
 Female:2543   Min.   :0.0000   Min.   :11.00   Min.   :  0.000
 Male  :3192   1st Qu.:0.4709   1st Qu.:16.06   1st Qu.:  8.398
               Median :0.6280   Median :19.75   Median : 14.100
               Mean   :0.5925   Mean   :20.50   Mean   : 15.645
               3rd Qu.:0.7430   3rd Qu.:23.43   3rd Qu.: 20.049
               Max.   :0.9620   Max.   :33.00   Max.   :192.000

     pafi1          swang1        wtkilo1       gastr       meta
 Min.   : 11.6   No RHC:3551   Min.   :  0.00   No :4793   No :5470
 1st Qu.:133.3   RHC   :2184   1st Qu.: 56.30   Yes: 942   Yes: 265
 Median :202.5                 Median : 70.00
 Mean   :222.3                 Mean   : 67.83
 3rd Qu.:316.6                 3rd Qu.: 83.70
 Max.   :937.5                 Max.   :244.00

 hema          adld3p         urin1
 No :5381   0      :5074   Min.   :   0
 Yes: 354   1      : 296   1st Qu.:1927
            2      : 130   Median :1927
            5      :  64   Mean   :2052
            6      :  53   3rd Qu.:1927
            4      :  49   Max.   :9000
            (Other):  69
```

1.2 Design: Propensity Score Estimation and Balance Check

We fit the propensity score model by logistic regression and visually assess the overlap.

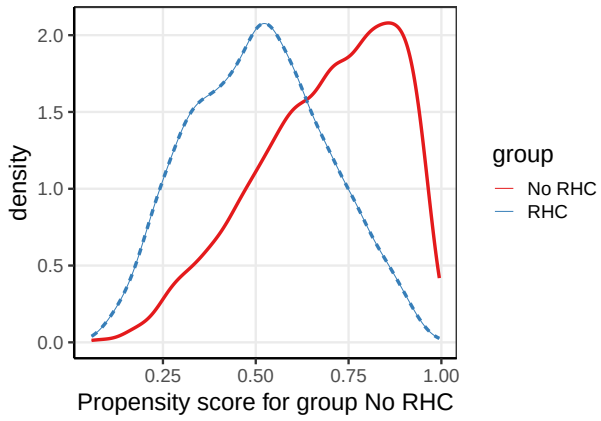
```
#Specify a logistic regression model: ps.rhc
ps.rhc <- swang1 ~ as.factor(cat1) + as.factor(cat2) + as.factor(ca) + pafi1 +
  wtkilo1 + surv2md1 + as.factor(dementhx) + as.factor(gastr) + wblc1 + temp1 +
   das2d3pc + age + as.factor(chfhx) + as.factor(sex) + urin1 +
  as.factor(hema) + as.factor(chrpulhx) + as.factor(cardiohx) +
  as.factor(meta) + as.factor(adld3p)
#Obtain the propensity score estimates
bal.rhc <- SumStat(ps.formula = ps.rhc, weight = c('IPW','overlap','treated'),
                   data = rhc)
summary(glm(ps.rhc,data=rhc,family = "binomial"))
```

```
Call:
glm(formula = ps.rhc, family = "binomial", data = rhc)

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                     2.500e+00  7.326e-01   3.412 0.000645 ***
as.factor(cat1)CHF              9.181e-01  1.460e-01   6.287 3.23e-10 ***
as.factor(cat1)Coma            -1.141e+00  1.600e-01  -7.131 9.96e-13 ***
as.factor(cat1)COPD            -1.061e+00  1.720e-01  -6.164 7.08e-10 ***
as.factor(cat1)MOSF             7.706e-01  8.275e-02   9.312  < 2e-16 ***
as.factor(cat1)Other           -9.594e-01  1.833e-01  -5.235 1.65e-07 ***
as.factor(cat2)Cirrhosis       -7.079e-01  3.877e-01  -1.826 0.067870 .
as.factor(cat2)Colon Cancer     1.510e+00  1.510e+00   1.000 0.317415
as.factor(cat2)Coma            -1.226e+00  2.890e-01  -4.244 2.19e-05 ***
as.factor(cat2)Lung Cancer     -2.831e-01  7.958e-01  -0.356 0.722065
as.factor(cat2)MOSF w/Malignancy -1.200e-01  1.949e-01  -0.615 0.538232
as.factor(cat2)MOSF w/Sepsis    6.054e-01  8.898e-02   6.804 1.01e-11 ***
as.factor(ca)No                 1.094e+00  1.536e-01   7.122 1.06e-12 ***
as.factor(ca)Yes                3.451e-01  1.515e-01   2.278 0.022708 *
pafi1                          -3.967e-03  3.096e-04 -12.811  < 2e-16 ***
wtkilo1                         7.668e-03  1.102e-03   6.960 3.40e-12 ***
surv2md1                       -2.216e+00  2.564e-01  -8.642  < 2e-16 ***
as.factor(dementhx)1           -6.633e-01  1.130e-01  -5.871 4.33e-09 ***
as.factor(gastr)Yes             3.284e-01  8.777e-02   3.741 0.000183 ***
wblc1                           2.606e-03  2.594e-03   1.005 0.314997
temp1                          -5.552e-02  1.812e-02  -3.064 0.002181 **
das2d3pc                        3.123e-04  6.393e-03   0.049 0.961033
age                            -7.037e-03  2.145e-03  -3.280 0.001037 **
as.factor(chfhx)1               1.677e-01  9.688e-02   1.731 0.083399 .
as.factor(sex)Male              4.691e-02  6.267e-02   0.749 0.454095
urin1                           2.838e-05  2.824e-05   1.005 0.314839
as.factor(hema)Yes             -6.336e-01  1.391e-01  -4.554 5.27e-06 ***
as.factor(chrpulhx)1           -1.244e-01  9.397e-02  -1.324 0.185472
as.factor(cardiohx)1            1.976e-01  8.953e-02   2.207 0.027330 *
as.factor(meta)Yes             -1.304e-01  1.466e-01  -0.890 0.373653
as.factor(adld3p)1             -5.729e-01  1.512e-01  -3.789 0.000151 ***
as.factor(adld3p)2             -4.373e-01  2.202e-01  -1.986 0.047082 *
as.factor(adld3p)3             -5.467e-01  3.764e-01  -1.452 0.146385
as.factor(adld3p)4             -8.317e-01  3.965e-01  -2.098 0.035925 *
```

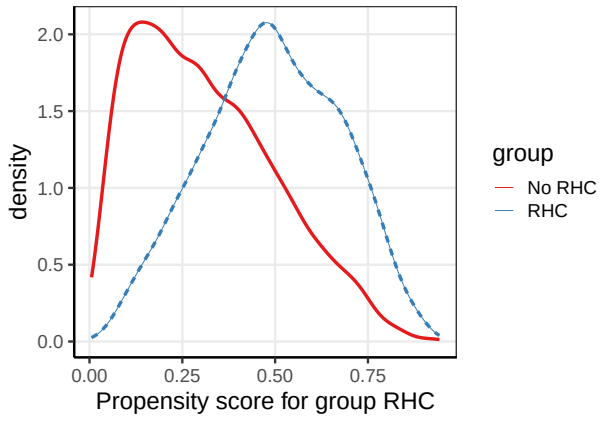
```
as.factor(adld3p)5               -4.247e-01  3.033e-01  -1.400 0.161367
as.factor(adld3p)6               -8.484e-01  3.682e-01  -2.304 0.021219 *
as.factor(adld3p)7               -6.693e-01  5.427e-01  -1.233 0.217441
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7621.4  on 5734  degrees of freedom
Residual deviance: 6514.3  on 5698  degrees of freedom
AIC: 6588.3

Number of Fisher Scoring iterations: 4
```

```r
#Density plots of estimated propensity scores
plot(bal.rhc, type = 'density')
```

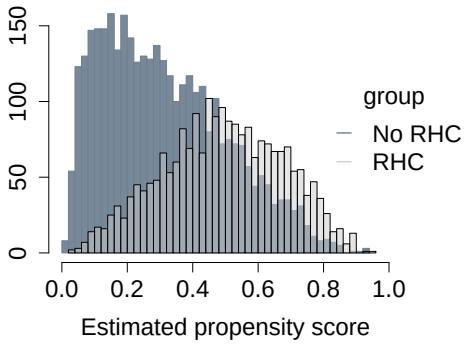
```
Propensity score for group No RHC
```




```
Press [enter] to continue

Propensity score for group RHC
```



```
plot(bal.rhc, type = 'hist')
```




```
summary(bal.rhc)
```

```
unweighted result
                                 Mean No RHC Mean RHC   SMD
as.factor(cat1)CHF                      0.070    0.096 0.095
as.factor(cat1)Coma                     0.096    0.043 0.207
as.factor(cat1)COPD                     0.112    0.027 0.342
as.factor(cat1)MOSF                     0.216    0.393 0.391
as.factor(cat1)Other                    0.061    0.025 0.175
as.factor(cat2)Cirrhosis                0.008    0.005 0.032
as.factor(cat2)Colon Cancer             0.000    0.000 0.009
as.factor(cat2)Coma                     0.020    0.009 0.089
as.factor(cat2)Lung Cancer              0.004    0.001 0.057
as.factor(cat2)MOSF w/Malignancy        0.048    0.027 0.114
as.factor(cat2)MOSF w/Sepsis            0.114    0.192 0.218
as.factor(ca)No                         0.747    0.791 0.104
as.factor(ca)Yes                        0.180    0.153 0.072
pafi1                                 240.627  192.433 0.433
wtkilo1                                65.040   72.360 0.256
surv2md1                                0.607    0.568 0.198
as.factor(dementhx)1                    0.116    0.069 0.163
as.factor(gastr)Yes                     0.147    0.192 0.121
wblc1                                  15.263   16.266 0.084
temp1                                  37.633   37.595 0.021
das2d3pc                               20.371   20.701 0.063
age                                    61.761   60.750 0.061
as.factor(chfhx)1                       0.168    0.195 0.069
as.factor(sex)Male                      0.539    0.585 0.093
urin1                                2049.945 2056.123 0.006
as.factor(hema)Yes                      0.067    0.053 0.062
as.factor(chrpulhx)1                    0.218    0.144 0.192
as.factor(cardiohx)1                    0.160    0.204 0.116
as.factor(meta)Yes                      0.048    0.043 0.028
as.factor(adld3p)1                      0.063    0.033 0.138
as.factor(adld3p)2                      0.027    0.016 0.079
as.factor(adld3p)3                      0.009    0.005 0.048
as.factor(adld3p)4                      0.011    0.004 0.082
as.factor(adld3p)5                      0.013    0.008 0.053
```

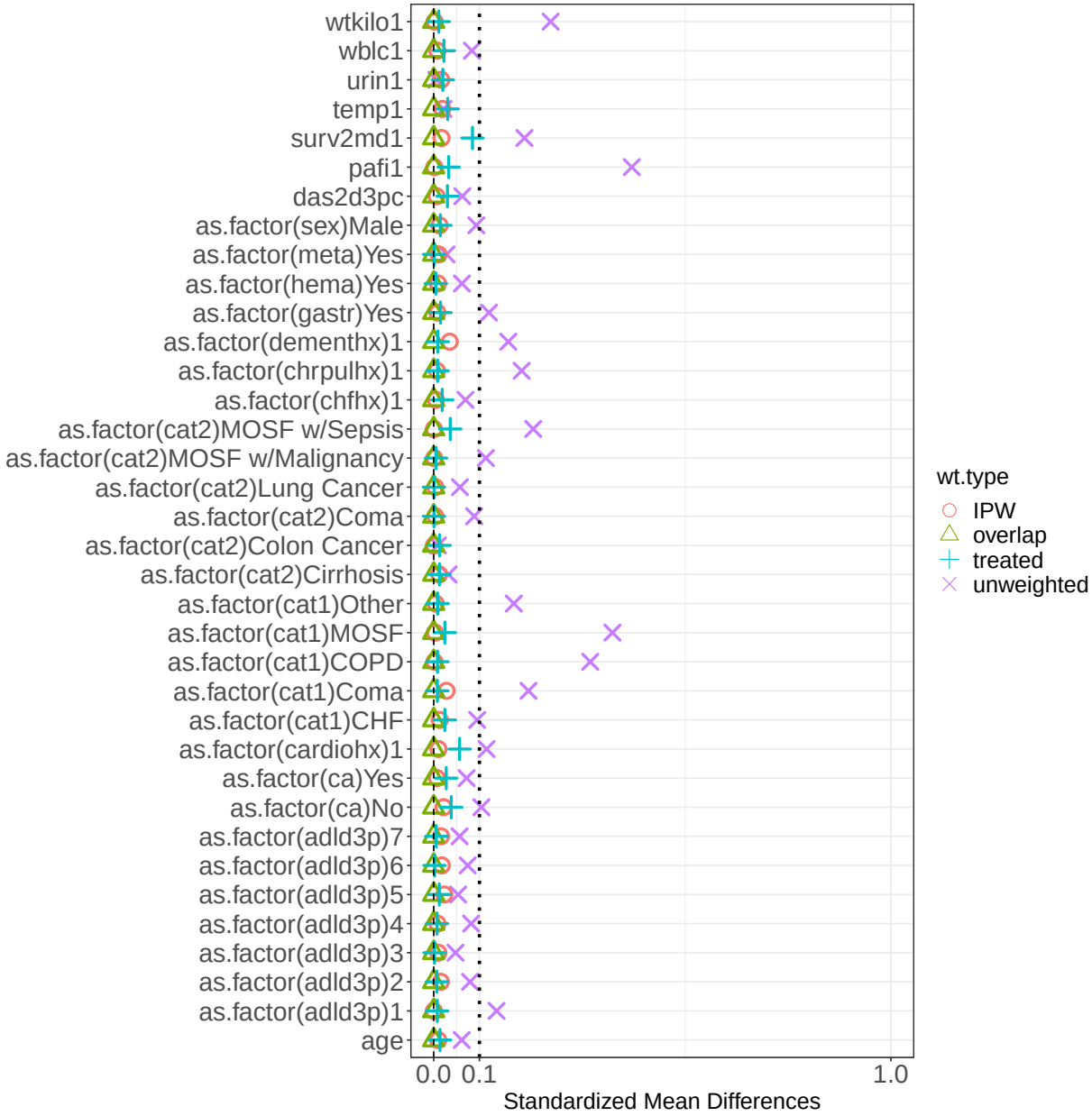
```
as.factor(adld3p)6                     0.012    0.005 0.074
as.factor(adld3p)7                     0.006    0.002 0.057

IPW result
                                 Mean No RHC Mean RHC   SMD
as.factor(cat1)CHF                     0.083    0.086 0.013
as.factor(cat1)Coma                    0.075    0.068 0.028
as.factor(cat1)COPD                    0.078    0.078 0.002
as.factor(cat1)MOSF                    0.290    0.288 0.003
as.factor(cat1)Other                   0.046    0.047 0.005
as.factor(cat2)Cirrhosis               0.006    0.005 0.012
as.factor(cat2)Colon Cancer            0.000    0.000 0.001
as.factor(cat2)Coma                    0.016    0.015 0.004
as.factor(cat2)Lung Cancer             0.003    0.002 0.004
as.factor(cat2)MOSF w/Malignancy       0.039    0.039 0.001
as.factor(cat2)MOSF w/Sepsis           0.150    0.150 0.000
as.factor(ca)No                        0.758    0.748 0.022
as.factor(ca)Yes                       0.173    0.176 0.007
pafi1                                220.558  220.698 0.001
wtkilo1                               68.008   68.067 0.002
surv2md1                               0.585    0.582 0.017
as.factor(dementhx)1                   0.097    0.087 0.035
as.factor(gastr)Yes                    0.167    0.164 0.008
wblc1                                 15.768   15.871 0.009
temp1                                 37.595   37.561 0.019
das2d3pc                              20.440   20.472 0.006
age                                   61.279   61.440 0.010
as.factor(chfhx)1                      0.181    0.183 0.005
as.factor(sex)Male                     0.554    0.560 0.013
urin1                               2061.024 2078.463 0.017
as.factor(hema)Yes                     0.062    0.064 0.010
as.factor(chrpulhx)1                   0.188    0.191 0.007
as.factor(cardiohx)1                   0.186    0.190 0.010
as.factor(meta)Yes                     0.046    0.044 0.011
as.factor(adld3p)1                     0.052    0.052 0.000
as.factor(adld3p)2                     0.022    0.020 0.016
as.factor(adld3p)3                     0.007    0.007 0.010
as.factor(adld3p)4                     0.008    0.008 0.008
as.factor(adld3p)5                     0.011    0.008 0.023
as.factor(adld3p)6                     0.009    0.011 0.019
as.factor(adld3p)7                     0.005    0.004 0.016

overlap result
                                 Mean No RHC Mean RHC SMD
as.factor(cat1)CHF                     0.097    0.097   0
as.factor(cat1)Coma                    0.057    0.057   0
as.factor(cat1)COPD                    0.043    0.043   0
as.factor(cat1)MOSF                    0.320    0.320   0
as.factor(cat1)Other                   0.035    0.035   0
as.factor(cat2)Cirrhosis               0.006    0.006   0
as.factor(cat2)Colon Cancer            0.000    0.000   0
as.factor(cat2)Coma                    0.013    0.013   0
as.factor(cat2)Lung Cancer             0.001    0.001   0
as.factor(cat2)MOSF w/Malignancy       0.035    0.035   0
```

```
as.factor(cat2)MOSF w/Sepsis           0.163    0.163   0
as.factor(ca)No                        0.771    0.771   0
as.factor(ca)Yes                       0.167    0.167   0
pafi1                                209.566  209.566   0
wtkilo1                               69.752   69.752   0
surv2md1                               0.585    0.585   0
as.factor(dementhx)1                   0.082    0.082   0
as.factor(gastr)Yes                    0.170    0.170   0
wblc1                                 15.845   15.845   0
temp1                                 37.624   37.624   0
das2d3pc                              20.639   20.639   0
age                                   60.925   60.925   0
as.factor(chfhx)1                      0.192    0.192   0
as.factor(sex)Male                     0.570    0.570   0
urin1                              2077.393 2077.393   0
as.factor(hema)Yes                     0.060    0.060   0
as.factor(chrpulhx)1                   0.160    0.160   0
as.factor(cardiohx)1                   0.199    0.199   0
as.factor(meta)Yes                     0.045    0.045   0
as.factor(adld3p)1                     0.044    0.044   0
as.factor(adld3p)2                     0.019    0.019   0
as.factor(adld3p)3                     0.006    0.006   0
as.factor(adld3p)4                     0.006    0.006   0
as.factor(adld3p)5                     0.009    0.009   0
as.factor(adld3p)6                     0.007    0.007   0
as.factor(adld3p)7                     0.003    0.003   0

treated result
                                Mean No RHC Mean RHC   SMD
as.factor(cat1)CHF                   0.103    0.096 0.026
as.factor(cat1)Coma                  0.042    0.043 0.008
as.factor(cat1)COPD                  0.025    0.027 0.008
as.factor(cat1)MOSF                  0.405    0.393 0.026
as.factor(cat1)Other                 0.024    0.025 0.009
as.factor(cat2)Cirrhosis             0.004    0.005 0.013
as.factor(cat2)Colon Cancer          0.000    0.000 0.013
as.factor(cat2)Coma                  0.009    0.009 0.001
as.factor(cat2)Lung Cancer           0.001    0.001 0.001
as.factor(cat2)MOSF w/Malignancy     0.026    0.027 0.005
as.factor(cat2)MOSF w/Sepsis         0.207    0.192 0.040
as.factor(ca)No                      0.775    0.791 0.038
as.factor(ca)Yes                     0.163    0.153 0.027
pafi1                              189.029  192.433 0.033
wtkilo1                             72.671   72.360 0.011
surv2md1                             0.551    0.568 0.090
as.factor(dementhx)1                 0.067    0.069 0.009
as.factor(gastr)Yes                  0.198    0.192 0.016
wblc1                               16.560   16.266 0.025
temp1                               37.536   37.595 0.033
das2d3pc                            20.547   20.701 0.030
age                                 60.523   60.750 0.014
as.factor(chfhx)1                    0.202    0.195 0.019
as.factor(sex)Male                   0.578    0.585 0.015
urin1                             2078.429 2056.123 0.021
```

```
as.factor(hema)Yes                  0.054    0.053 0.005
as.factor(chrpulhx)1                0.141    0.144 0.009
as.factor(cardiohx)1                0.228    0.204 0.060
as.factor(meta)Yes                  0.043    0.043 0.001
as.factor(adld3p)1                  0.035    0.033 0.008
as.factor(adld3p)2                  0.015    0.016 0.007
as.factor(adld3p)3                  0.005    0.005 0.002
as.factor(adld3p)4                  0.004    0.004 0.007
as.factor(adld3p)5                  0.007    0.008 0.012
as.factor(adld3p)6                  0.005    0.005 0.002
as.factor(adld3p)7                  0.003    0.002 0.006
```

Overall, the overlap of this data set is satisfactory by checking the density plots (in this case, we would expect the true causal effects under different weighting schemes to be similar, but OW tends to have the smallest variance.)

```r
#Check covariates balance using the metric average standardized differences (ASD)
plot(bal.rhc, metric = 'ASD')
```



From the Love plot on the comparisons between original standardized mean differences (ASD) of all covariates and ASD of all covariates after IPW, OW, ATT weighting methods, we observe that the weighted ASD of all covariates are improved (under three different weighting methods) compared to the unweighted. In particular, the resulting weighted ASD for all covariates after OW are exactly 0.

1.3 Outcome Regression

The key step in outcome regression is to predict the missing counterfactual outcomes for each unit based on the fitted model. With these predicted outcomes, we can estimate any causal estimands.

We fit two logistic outcome models: (i) with the main effects of all covariates; (ii) with some main effects and some interactions between treatment and main effects, selected by LASSO (post-LASSO).

(i) Logisitic regression with all main effects

```
logis_reg_main_effects <- glm(
  death ~ as.factor(cat1) + as.factor(cat2) + as.factor(ca) + pafi1 +
    wtkilo1 + surv2md1 + as.factor(dementhx) + as.factor(gastr) + wblc1 + temp1 +
    das2d3pc + age + as.factor(chfhx) + as.factor(sex) + urin1 +
    as.factor(hema) + as.factor(chrpulhx) + as.factor(cardiohx) +
    as.factor(meta) + as.factor(adld3p) + swang1,
  family = binomial("logit"), data = rhc
)
```

(ii) Logistic regression with main effects and interactions (LASSO selection)

```
X_matrix <- model.matrix(
  death ~ (as.factor(cat1) + as.factor(cat2) + as.factor(ca) + pafi1 +
    wtkilo1 + surv2md1 + as.factor(dementhx) + as.factor(gastr) + wblc1 + temp1 +
    das2d3pc + age + as.factor(chfhx) + as.factor(sex) + urin1 +
    as.factor(hema) + as.factor(chrpulhx) + as.factor(cardiohx) +
    as.factor(meta) + as.factor(adld3p)) * swang1,
  family = binomial("logit"), data = rhc)

cv.glmmod <- cv.glmnet(X_matrix[, -1], rhc$death, alpha=1, family = binomial())
best.lambda <- cv.glmmod$lambda.min
glmmod <- glmnet(X_matrix[, -1], rhc$death, family = binomial(), lambda = best.lambda)

active.index <- which(coef(glmmod, s = best.lambda) != 0)
active_X <- X_matrix[, active.index]

logis_reg_LASSO <- glm(
  death ~ ., family = binomial('logit'),
  data = as.data.frame(cbind(active_X[, -1], death = rhc$death))
)
```

Now, we shall predict the probability of death of each unit under treatment and control, respectively. There is a subtlety here: we could either (i) only predict the missing potential outcomes or (ii) predict both potential outcomes for each unit. Theoretically, this is the difference between population- and sample- estimands. Here we use the former. Note that, with large sample, the numerical difference between the two is usually little.

```
rhc_trt <- rhc
rhc_trt$swang1 = 'RHC'

rhc_ctrl <- rhc
rhc_ctrl$swang1 = 'No RHC'

pred_prob_main_effects_trt <- predict(logis_reg_main_effects, rhc_trt, type = 'response')
pred_prob_main_effects_ctrl <- predict(logis_reg_main_effects, rhc_ctrl, type = 'response')

X_matrix_trt <- X_matrix
X_matrix_trt[, 38] = 1
X_matrix_trt[, 39:74] = X_matrix_trt[, 2:37]

X_matrix_ctrl <- X_matrix
X_matrix_ctrl[, 38:74] = 0

pred_prob_LASSO_trt <- predict(
  logis_reg_LASSO, as.data.frame(X_matrix_trt), type = 'response')
pred_prob_LASSO_ctrl <- predict(
```

```
  logis_reg_LASSO, as.data.frame(X_matrix_ctrl), type = 'response')
```

Because we have binary outcome, the next step is to draw the potential outcomes for each unit based on a Bernoulli with the predicted probabilities. Given these draws, we can calculate the RD, RR and OR.

```
pred_out_regression <- function(prob_trt, prob_ctrl, subset = TRUE,
                                weights = rep(1, length(prob_trt[subset]))) {
  out_trt <- rbinom(length(prob_trt), 1, prob_trt)[subset]
  out_ctrl <- rbinom(length(prob_ctrl), 1, prob_ctrl)[subset]
  RD <- weighted.mean(out_trt, weights) - weighted.mean(out_ctrl, weights)
  RR <- weighted.mean(out_trt) / weighted.mean(out_ctrl)
  OR <- (weighted.mean(out_trt) / (1 - weighted.mean(out_trt))) /
    (weighted.mean(out_ctrl) / (1 - weighted.mean(out_ctrl)))
  return (c(RD = RD, RR = RR, OR = OR))
}
```

We can obtain the confidence intervals via bootstrap.

```
bootstrap <- function(func, num_rep, ...) {
  set.seed(0)
  pb <- progress_bar$new(
    format = "[:bar] :current/:total  elapsed: :elapsedfull  eta: :eta",
    total = num_rep, clear = TRUE, width = 60, show_after = 10
  )
  bootstrap_results = list()
  for (i in 1:num_rep) {
    bootstrap_results[[i]] <- func(...)
    pb$tick()
  }
  res <- do.call(rbind, bootstrap_results)
  return (apply(res, 2, function(x) c(mean(x), quantile(x, 0.025), quantile(x, 0.975))))
}
```

1.3.1 ATE Estimand

```
ATE_main_effects <- bootstrap(pred_out_regression, 1000,
                              prob_trt = pred_prob_main_effects_trt,
                              prob_ctrl = pred_prob_main_effects_ctrl)
ATE_LASSO <- bootstrap(pred_out_regression, 1000,
                       prob_trt = pred_prob_LASSO_trt,
                       prob_ctrl = pred_prob_LASSO_ctrl)
```

```
ATE_main_effects
```

```
              RD       RR       OR
      0.04561813 1.072289 1.224691
2.5%  0.02859634 1.044371 1.135895
97.5% 0.06137751 1.098426 1.311156
```

```
ATE_LASSO
```

```
              RD       RR       OR
      0.04889730 1.077226 1.244079
2.5%  0.03225371 1.050116 1.154603
97.5% 0.06451613 1.103197 1.332112
```

The regression results using main effects and post-LASSO effects are similar. Using main effects, the risk difference is 0.046 (0.029 to 0.061), the relative risk is 1.072 (1.044 to 1.098) and the odds ratio is 1.225 (1.136 to 1.311). Using post-LASSO effects, the risk difference is 0.049 (0.033 to 0.065), the relative risk is 1.078 (1.050 to 1.104) and the odds ratio is 1.245 (1.157 to 1.334).

1.3.2 ATT estimand

```
ATT_main_effects <- bootstrap(pred_out_regression, 1000,
                              prob_trt = pred_prob_main_effects_trt,
                              prob_ctrl = pred_prob_main_effects_ctrl,
                              subset = rhc$swang1 == 'RHC')
ATT_LASSO <- bootstrap(pred_out_regression, 1000,
                       prob_trt = pred_prob_LASSO_trt,
                       prob_ctrl = pred_prob_LASSO_ctrl,
                       subset = rhc$swang1 == 'RHC')
```

```
ATT_main_effects
```

```
              RD       RR       OR
      0.04556227 1.071936 1.226772
2.5%  0.02060440 1.032278 1.096860
97.5% 0.07052427 1.113919 1.371710
```

```
ATT_LASSO
```

```
              RD       RR       OR
      0.04013049 1.062909 1.198010
2.5%  0.01648352 1.025028 1.076204
97.5% 0.06456044 1.102772 1.334128
```

The ATT estimates are also similar between the two models. Using the main effects model, the risk difference is 0.046 (0.021 to 0.070), the relative risk is 1.072 (1.032 to 1.114) and the odds ratio is 1.227 (1.097 to 1.372). Using post-LASSO effects, the risk difference is 0.040 (0.016 to 0.064), the relative risk is 1.063 (1.024 to 1.102) and the odds ratio is 1.198 (1.075 to 1.330).

1.4 Matching

We use the Matching R package to conduct matching. The default distance metric is the normalized Euclidean distance.

```
matching_matrix <- model.matrix(
  death ~ as.factor(cat1) + as.factor(cat2) + as.factor(ca) + pafi1 +
    wtkilo1 + surv2md1 + as.factor(dementhx) + as.factor(gastr) + wblc1 + temp1 +
    das2d3pc + age + as.factor(chfhx) + as.factor(sex) + urin1 +
    as.factor(hema) + as.factor(chrpulhx) + as.factor(cardiohx) +
    as.factor(meta) + as.factor(adld3p), data = rhc
)
match_obj_ATE <- Match(rhc_rep$death, ifelse(rhc$swang1 == "RHC", 1, 0),
                       X = matching_matrix, estimand = "ATE")
match_obj_ATT <- Match(rhc_rep$death, ifelse(rhc$swang1 == "RHC", 1, 0),
                       X = matching_matrix, estimand = "ATT")
matching_ATE <- pred_out_regression(rhc_rep$death[match_obj_ATE$index.treated],
                                    rhc_rep$death[match_obj_ATE$index.control])
```

```
matching_ATT <- pred_out_regression(rhc_rep$death[match_obj_ATT$index.treated],
                                    rhc_rep$death[match_obj_ATT$index.control])
```

```
matching_ATE
```

```
        RD         RR         OR
0.05632084 1.08895621 1.28644799
```

```
matching_ATT
```

```
        RD         RR         OR
0.04166667 1.06523297 1.20411005
```

The ATE estimates are: risk difference 0.056, relative risk 1.089, odds ratio 1.286. The ATT estimates are: risk difference 0.042, relative risk 1.065, odds ratio 1.204.

Note that Abadie and Imbens (2008) showed that bootstrap leads to biased variance estimate and Lin et al. (2023) provided a simple closed-form fix (see Ding's textbook page 207-208). Nevertheless, here we illustrate the use of bootstrap: (1) resample the data, (2)obtain new matches and causal estimates using the bootstrap sample, (3) repeat step 1-2.

```
bootstrap_matching <- function(estimand) {
  indices <- sample(1:nrow(rhc), nrow(rhc), replace = TRUE)
  outcome <- rhc_rep$death[indices]
  treatment <- ifelse(rhc$swang1 == "RHC", 1, 0)
  match_obj <- Match(Y = outcome, Tr = treatment[indices],
                     X = matching_matrix[indices, ], estimand = estimand)
  return (pred_out_regression(
    outcome[match_obj$index.treated], outcome[match_obj$index.control],
    weights = match_obj$weights
  ))
}
```

```
matching_bootstrap_ATE <- bootstrap(bootstrap_matching, num_rep = 100, estimand = "ATE")
matching_bootstrap_ATT <- bootstrap(bootstrap_matching, num_rep = 100, estimand = "ATT")
```

```
matching_bootstrap_ATE
```

```
              RD       RR      OR
      0.05916071 1.093908 1.30465
2.5%  0.03390148 1.044057 1.12710
97.5% 0.08475153 1.143774 1.49505
```

The confidence intervals for ATE are: [0.034, 0.085] for risk difference, [1.044, 1.144] for relative risk, and [1.127, 1.495] for odds ratio.

```
matching_bootstrap_ATT
```

```
              RD       RR       OR
      0.05025300 1.080478 1.252488
2.5%  0.01721734 1.021515 1.065249
97.5% 0.07809065 1.156493 1.488035
```

The confidence intervals for ATT are: [0.017, 0.078] for risk difference, [1.022, 1.156] for relative risk, and [1.065, 1.488] for odds ratio.

1.5 Weighting

In this section, we demonstrate several weighting methods including inverse probability weighting, overlap weighting, and ATT weighting.

Our first estimand is the causal risk difference among the combined population (ATE). We use IPW weighting method first and compare the standard errors estimated by robust sandwich estimator (default choice) or bootstrap.

```
#Implement IPW weighting method and estimating standard errors by robust
set.seed(0)
#sandwich estimator
ate.rhc.ipw <- PSweight(ps.formula = ps.rhc, yname = "death", data = rhc,
                        weight = 'IPW')
#Implement IPW weighting method and estimating standard errors by bootstrap
ate.rhc.ipw.boot <- PSweight(ps.formula = ps.rhc, yname = "death", data = rhc,
                        weight = 'IPW', bootstrap = TRUE)
```

```
bootstrap 50 samples
```

```
#Target estimand is the contrast between treatment group and control group
contrast.rhc <- c(1,-1)
#Check the format of output
summary(ate.rhc.ipw, type = 'DIF', constrast=contrast.rhc)
```

```
Closed-form inference:

Original group value:  No RHC, RHC

Contrast:
           No RHC RHC
Contrast 1     -1   1

           Estimate Std.Error      lwr     upr Pr(>|z|)
Contrast 1 0.042960  0.013811 0.015890 0.07003 0.001868 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Causal risk difference estimates
summary(ate.rhc.ipw,
        type = 'DIF', constrast=contrast.rhc)$estimates[,c(1,2,4,5)]
```

```
  Estimate  Std.Error        lwr        upr
0.04295983 0.01381139 0.01589000 0.07002966
```

```
summary(ate.rhc.ipw.boot,
        type = 'DIF', constrast=contrast.rhc)$estimates[,c(1,2,4,5)]
```

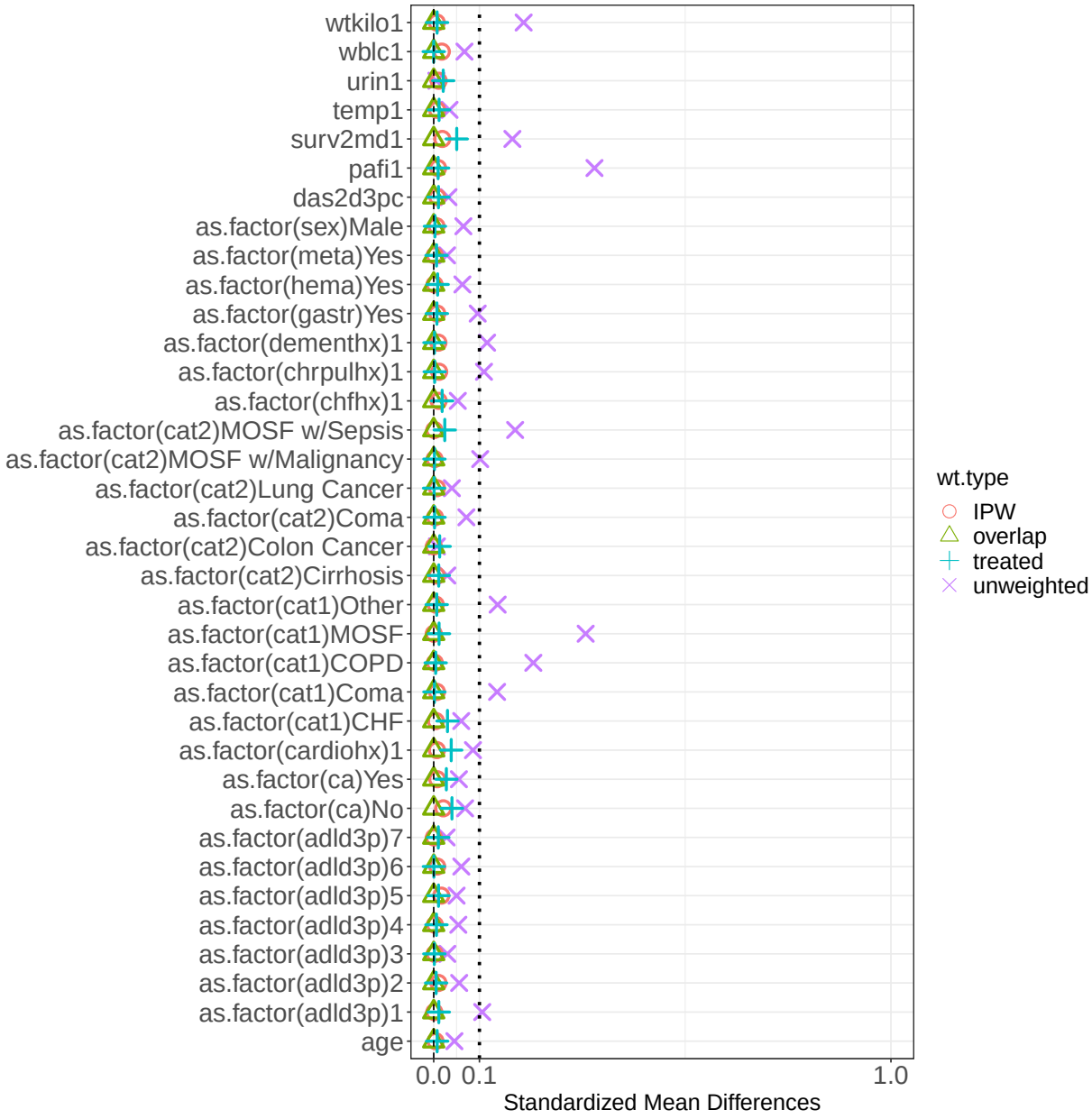
```
  Estimate  Std.Error        lwr        upr
0.04295983 0.01074851 0.02329148 0.05910196
```

We note that the estimates of standard error by robust sandwich estimator and bootstrap are very close in this example. Hereafter, we consider the robust sandwich estimator of the standard error as the computational cost of robust sandwich estimator is much less.

The causal risk difference estimate using IPW weighting method is 0.043. The corresponding standard error estimate is 0.014. The 95% confidence interval is $[0.016, 0.070]$.

Next, we consider to estimate using IPW trimming method. We use the rule of thumb threshold $\delta = 0.1$, i.e., excluding those whose estimated propensity scores are less than 0.1 or bigger than 0.9.

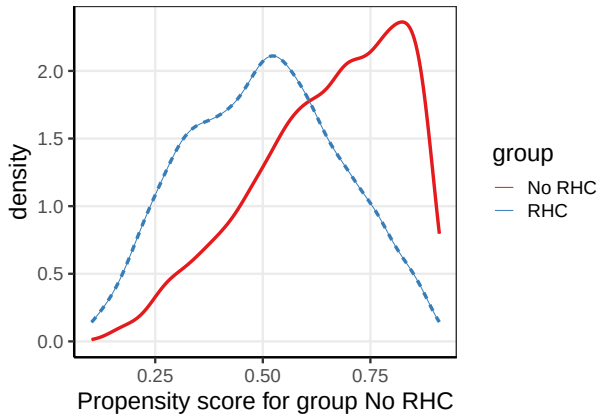
```
#Apply trimming with delta=0.1
bal.rhc.trim <- SumStat(ps.formula = ps.rhc,
    weight = c('IPW','overlap','treated'),
    data = rhc, delta = 0.1)
#Covariates balance after trimming
plot(bal.rhc.trim, metric = 'ASD')
```



We can see that the covariates balance after trimming is preserved and very similar to that without trimming.

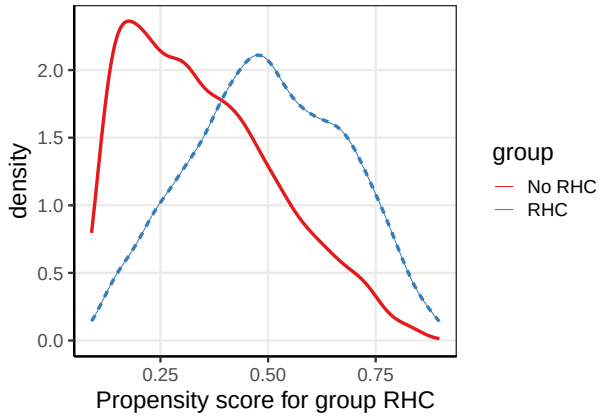
```
#Check the overlap
#Density plots of estimated propensity scores after trimming
plot(bal.rhc.trim, type = 'density')
```

```
Propensity score for group No RHC
```




```
Press [enter] to continue
```

```
Propensity score for group RHC
```



Overlap is still maintained as expected.

```
#Check how many patients are excluded
bal.rhc.trim$trim
```

```
          No RHC  RHC
trimmed      466   28
remained    3085 2156
```

We can see that nearly 8.8% patients are trimmed and they are mostly from the control group. We then estimate the causal risk difference using IPW after trimming.

```
#Implement IPW weighting method after trimming
ate.rhc.ipw.trim <- PSweight(ps.formula = ps.rhc, yname = "death", data = rhc,
                         weight = 'IPW', delta=0.1)
#Causal risk difference estimates
summary(ate.rhc.ipw.trim,
        type = 'DIF', constrast=contrast.rhc)$estimates[,c(1,2,4,5)]
```

```
  Estimate  Std.Error        lwr        upr
0.04074221 0.01317916 0.01491153 0.06657288
```

The causal risk difference estimate using IPW weighting method after trimming is 0.041. The corresponding

standard error estimate is 0.013. The 95% confidence interval is $[0.015, 0.067]$.

Next, we use ATT weighting method.

```
#Implement ATT weighting method
ate.rhc.att <- PSweight(ps.formula = ps.rhc, yname = "death", data = rhc,
                        weight = 'treated')
#Target estimand is the contrast between treatment group and control group
contrast.rhc <- c(1,-1)
#Causal risk difference estimates
summary(ate.rhc.att,
        type = 'DIF', constrast=contrast.rhc)$estimates[,c(1,2,4,5)]
```

```
     Estimate     Std.Error           lwr           upr
 0.0281025551  0.0145856796 -0.0004848516  0.0566899619
```

The causal risk difference estimate using ATT weighting method is 0.028. The corresponding standard error estimate is 0.015. The 95% confidence interval is $[-0.000, 0.057]$.

Next, we use OW weighting method.

```
#Implement IPW weighting method
ate.rhc.ow <- PSweight(ps.formula = ps.rhc, yname = "death", data = rhc,
                       weight = 'overlap')
#Target estimand is the contrast between treatment group and control group
contrast.rhc <- c(1,-1)
#Causal risk difference estimates
summary(ate.rhc.ow,
        type = 'DIF', constrast=contrast.rhc)$estimates[,c(1,2,4,5)]
```

```
  Estimate  Std.Error        lwr        upr
0.04438220 0.01264989 0.01958887 0.06917553
```

The causal risk difference estimate using OW weighting method is 0.044. The corresponding standard error estimate is 0.013. The 95% confidence interval is $[0.020, 0.069]$.

Second, we estimate the causal risk ratios using three weighting methods.

IPW:

```
#Causal risk ratio estimates
#standard error estimate
summary(ate.rhc.ipw, type = 'RR', constrast=contrast.rhc)$estimates[,2]
```

```
[1] 0.02072606
```

```
#point estimates are in the log scale, take exponential to go back RR scale
exp(summary(ate.rhc.ipw,
            type = 'RR', constrast=contrast.rhc)$estimates[,c(1,4,5)])
```

```
Estimate      lwr      upr
1.067285 1.024798 1.111533
```

The causal risk ratio estimate using IPW weighting method is 1.067. The corresponding standard error estimate is 0.021. The 95% confidence interval is $[1.025, 1.112]$.

ATT:

```
#Causal risk ratio estimates
#standard error estimate
summary(ate.rhc.att, type = 'RR', constrast=contrast.rhc)$estimates[,2]
```

```
[1] 0.02202241
```

```
#point estimates are in the log scale, take exponential to go back RR scale
exp(summary(ate.rhc.att,
            type = 'RR', constrast=contrast.rhc)$estimates[,c(1,4,5)])
```

```
 Estimate       lwr       upr
1.0430822 0.9990174 1.0890907
```

The causal risk ratio estimate using ATT weighting method is 1.043. The corresponding standard error estimate is 0.022. The 95% confidence interval is $[0.999, 1.089]$.

OW:

```
#Causal risk ratio estimates
#standard error estimate
summary(ate.rhc.ow, type = 'RR', constrast=contrast.rhc)$estimates[,2]
```

```
[1] 0.01928697
```

```
#point estimates are in the log scale, take exponential to go back RR scale
exp(summary(ate.rhc.ow,
            type = 'RR', constrast=contrast.rhc)$estimates[,c(1,4,5)])
```

```
Estimate      lwr      upr
1.070217 1.030516 1.111447
```

The causal risk ratio estimate using OW weighting method is 1.070. The corresponding standard error estimate is 0.019. The 95% confidence interval is $[1.031, 1.111]$.

Third, we estimate the causal odds ratio using three weighting methods.

IPW:

```
#Causal odds ratio estimates
#standard error estimate
summary(ate.rhc.ipw, type = 'OR', constrast=contrast.rhc)$estimates[,2]
```

```
[1] 0.0624144
```

```
#point estimates are in the log scale, take exponential to go back RR scale
exp(summary(ate.rhc.ipw,
            type = 'OR', constrast=contrast.rhc)$estimates[,c(1,4,5)])
```

```
Estimate      lwr      upr
1.211214 1.071751 1.368826
```

The causal odds ratio estimate using IPW weighting method is 1.211. The corresponding standard error estimate is 0.062. The 95% confidence interval is $[1.072, 1.369]$.

ATT:

```
#Causal odds ratio estimates
#standard error estimate
summary(ate.rhc.att, type = 'OR', constrast=contrast.rhc)$estimates[,2]
```

```
[1] 0.06535963
```

```
#point estimates are in the log scale, take exponential to go back RR scale
exp(summary(ate.rhc.att,
            type = 'OR', constrast=contrast.rhc)$estimates[,c(1,4,5)])
```

```
 Estimate       lwr       upr
```

```
1.1348017 0.9983568 1.2898945
```

The causal odds ratio estimate using ATT weighting method is 1.135. The corresponding standard error estimate is 0.065. The 95% confidence interval is $[0.998, 1.290]$.

OW:

```
#Causal odds ratio estimates
#standard error estimate
summary(ate.rhc.ow, type = 'OR', constrast=contrast.rhc)$estimates[,2]
```

```
[1] 0.05634626
```

```
#point estimates are in the log scale, take exponential to go back RR scale
exp(summary(ate.rhc.ow,
            type = 'OR', constrast=contrast.rhc)$estimates[,c(1,4,5)])
```

```
Estimate      lwr      upr
1.217024 1.089776 1.359131
```

The causal odds ratio estimate using OW weighting method is 1.217. The corresponding standard error estimate is 0.056. The 95% confidence interval is $[1.090, 1.359]$.

Overall, we can see that OW weighting method has the least standard errors or the most narrow confidence intervals for the all target estimands.

1.6 Augmented Weighting

In this section, we introduce the augmented weighting method and follow the same procedure as the previous section. For IPW, augmented weighting is also known as the doubly-robust estimation. It is only singly-robust for other weighting schemes. We firstly estimate the causal risk difference.

Using the augmented IPW method:

```
#Specify a logistic regression model for the outcome
out.rhc <- death ~ as.factor(cat1) + as.factor(cat2) + as.factor(ca) + pafi1 +
  wtkilo1 + surv2md1 + as.factor(dementhx) + as.factor(gastr) + wblc1 + temp1 +
   das2d3pc + age + as.factor(chfhx) + as.factor(sex) + urin1 +
  as.factor(hema) + as.factor(chrpulhx) + as.factor(cardiohx) +
  as.factor(meta) + as.factor(adld3p)
#Implement augmented IPW weighting method
ate.rhc.ipw.aug <- PSweight(ps.formula = ps.rhc, yname = "death", data = rhc,
                            out.formula = out.rhc, weight = 'IPW',
                            augmentation = TRUE, family = "binomial")
```

```
Warning in value[[3L]](cond): The sandwich matrix not pd, therefore not
invertable, use conservative variance instead, please double check
```

```
#Target estimand is the contrast between treatment group and control group
contrast.rhc <- c(1,-1)
#Causal risk difference estimates
summary(ate.rhc.ipw.aug, type = 'DIF',
        constrast=contrast.rhc)$estimates[,c(1,2,4,5)]
```

```
  Estimate  Std.Error        lwr        upr
0.04284742 0.01377269 0.01585344 0.06984140
```

The causal risk difference estimate using the augmented IPW weighting method is 0.043. The corresponding standard error estimate is 0.014. The 95% confidence interval is $[0.016, 0.070]$.

Using the augmented ATT method:

```
#Implement augmented ATT weighting method
ate.rhc.att.aug <- PSweight(ps.formula = ps.rhc, yname = "death", data = rhc,
                            out.formula = out.rhc, weight = 'treated',
                            augmentation = TRUE, family = "binomial")
```

```
Warning in value[[3L]](cond): The sandwich matrix not pd, therefore not
invertable, use conservative variance instead, please double check
```

```
#Target estimand is the contrast between treatment group and control group
contrast.rhc <- c(1,-1)
#Causal risk difference estimates
summary(ate.rhc.att.aug, type = 'DIF',
        constrast=contrast.rhc)$estimates[,c(1,2,4,5)]
```

```
  Estimate  Std.Error        lwr        upr
0.04339651 0.01438657 0.01519935 0.07159368
```

The causal risk difference estimate using the augmented ATT weighting method is 0.043. The corresponding standard error estimate is 0.014. The 95% confidence interval is $[0.015, 0.072]$.

Using the augmented OW method:

```
#Implement augmented IPW weighting method
ate.rhc.ow.aug <- PSweight(ps.formula = ps.rhc, yname = "death", data = rhc,
                           out.formula = out.rhc, weight = 'overlap',
                           augmentation = TRUE, family = "binomial")
```

```
Warning in value[[3L]](cond): The sandwich matrix not pd, therefore not
invertable, use conservative variance instead, please double check
```

```
#Target estimand is the contrast between treatment group and control group
contrast.rhc <- c(1,-1)
#Causal risk difference estimates
summary(ate.rhc.ow.aug, type = 'DIF',
        constrast=contrast.rhc)$estimates[,c(1,2,4,5)]
```

```
  Estimate  Std.Error        lwr        upr
0.04720605 0.01253055 0.02264663 0.07176547
```

The causal risk difference estimate using the augmented OW weighting method is 0.047. The corresponding standard error estimate is 0.013. The 95% confidence interval is $[0.023, 0.072]$.

Second, we estimate the causal risk ratio.

Using the augmented IPW method:

```
#Causal risk ratio estimates
#standard error estimate
summary(ate.rhc.ipw.aug, type = 'RR', constrast=contrast.rhc)$estimates[,2]
```

```
[1] 0.02085751
```

```
#point estimates are in the log scale, take exponential to go back RR scale
exp(summary(ate.rhc.ipw.aug, type = 'RR',
            constrast=contrast.rhc)$estimates[,c(1,4,5)])
```

```
Estimate      lwr      upr
1.067720 1.024952 1.112273
```

The causal risk ratio estimate using IPW weighting method is 1.068. The corresponding standard error estimate is 0.021. The 95% confidence interval is $[1.025, 1.112]$.

Using the augmented ATT method:

```
#Causal risk ratio estimates
#standard error estimate
summary(ate.rhc.att.aug, type = 'RR', constrast=contrast.rhc)$estimates[,2]
```

```
[1] 0.02201632
```

```
#point estimates are in the log scale, take exponential to go back RR scale
exp(summary(ate.rhc.att.aug, type = 'RR',
            constrast=contrast.rhc)$estimates[,c(1,4,5)])
```

```
Estimate      lwr      upr
1.068016 1.022910 1.115111
```

The causal risk ratio estimate using the augmented ATT weighting method is 1.068. The corresponding standard error estimate is 0.022. The 95% confidence interval is $[1.023, 1.115]$.

Using the augmented OW method:

```
#Causal risk ratio estimates
#standard error estimate
summary(ate.rhc.ow.aug, type = 'RR', constrast=contrast.rhc)$estimates[,2]
```

```
[1] 0.01921291
```

```
#point estimates are in the log scale, take exponential to go back RR scale
exp(summary(ate.rhc.ow.aug, type = 'RR',
            constrast=contrast.rhc)$estimates[,c(1,4,5)])
```

```
Estimate      lwr      upr
1.075296 1.035557 1.116560
```

The causal risk ratio estimate using the augmented OW weighting method is 1.075. The corresponding standard error estimate is 0.019. The 95% confidence interval is $[1.036, 1.117]$.

Third, we estimate the causal odds ratio.

By the augmented IPW method:

```
#Causal odds ratio estimates
#standard error estimate
summary(ate.rhc.ipw.aug, type = 'OR', constrast=contrast.rhc)$estimates[,2]
```

```
[1] 0.06167894
```

```
#point estimates are in the log scale, take exponential to go back RR scale
exp(summary(ate.rhc.ipw.aug, type = 'OR',
            constrast=contrast.rhc)$estimates[,c(1,4,5)])
```

```
Estimate      lwr      upr
1.208729 1.071095 1.364050
```

The causal odds ratio estimate using IPW weighting method is 1.209. The corresponding standard error estimate is 0.062. The 95% confidence interval is $[1.071, 1.364]$.

By the augmented ATT method:

```r
#Causal odds ratio estimates
#standard error estimate
summary(ate.rhc.att.aug, type = 'OR', constrast=contrast.rhc)$estimates[,2]
```

```
[1] 0.06383974
```

```r
#point estimates are in the log scale, take exponential to go back RR scale
exp(summary(ate.rhc.att.aug, type = 'OR',
            constrast=contrast.rhc)$estimates[,c(1,4,5)])
```

```
Estimate      lwr      upr
1.213504 1.070781 1.375250
```

The causal odds ratio estimate using the augmented ATT weighting method is 1.214. The corresponding standard error estimate is 0.064. The 95% confidence interval is $[1.071, 1.375]$.

Using the agumented OW method:

```r
#Causal odds ratio estimates
#standard error estimate
summary(ate.rhc.ow.aug, type = 'OR', constrast=contrast.rhc)$estimates[,2]
```

```
[1] 0.05556381
```

```r
#point estimates are in the log scale, take exponential to go back RR scale
exp(summary(ate.rhc.ow.aug, type = 'OR',
            constrast=contrast.rhc)$estimates[,c(1,4,5)])
```

```
Estimate      lwr      upr
1.231073 1.104048 1.372713
```

The causal odds ratio estimate using the augmented OW weighting method is 1.231. The corresponding standard error estimate is 0.056. The 95% confidence interval is $[1.104, 1.373]$.

Overall, we can see that OW augmented weighting method has the smallest standard errors or the narrowest confidence intervals compared to other weighting methods.