Homework 4 - Kyungjin Cho

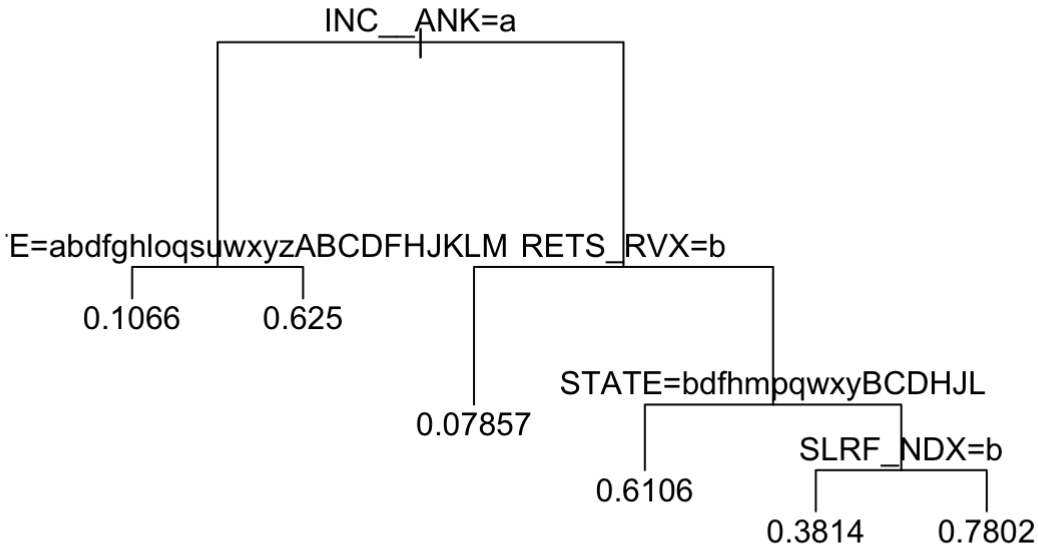
Data preparation

```
z <- read.table("/Users/irenecho/Desktop/stat443/datafiles/cedata.txt",header=TRUE)
w <- z$FINLWT21  ### sampling weights


zclass <- read.table("/Users/irenecho/Desktop/stat443/datafiles/ceclassid.fit",header=TR
UE)
probmissing <- zclass[,5] ### estimated P(INTRDVX_ = C)
p <- 1-probmissing
group <- !is.na(z$INTRDVX)
ipw <- sum(w[group]*z$INTRDVX[group]/p[group])/sum(w[group]/p[group])

zreg <- read.table("/Users/irenecho/Desktop/stat443/datafiles/ceregid.fit",header=TRUE)
yhat <- zreg$predicted
imputed <- (sum(w[group]*z$INTRDVX[group])+sum(w[!group]*yhat[!group]))/sum(w)
simple <- sum(w[group]*z$INTRDVX[group])/sum(w[group])
```

```
library(rpart)
vartype <- rep("numeric",638)
vartype[c(1,2,4,6,8,10,12,14,16,18,20,22,23,24,25,26,27,29,31,33,34,35,37,39,41,43,45,47
,49,52,54,56,58,60,62,64,66,68,70,72,73,74,75,76,77,78,79,80,82,84,86,88,89,90,92,94,96,
97,98,99,100,102,104,106,108,109,110,111,112,113,114,115,116,118,119,120,122,123,124,125
,126,128,130,131,132,133,135,137,139,303,304,305,306,307,308,309,310,311,312,313,314,315
,316,317,318,319,321,323,325,331,333,407,409,410,411,453,454,456,458,460,462,464,465,466
,467,468,470,472,474,476,477,478,479,482,484,486,488,490,492,494,496,497,498,499,500,502
,504,506,508,510,512,514,516,518,520,522,524,526,528,530,532,534,536,538,540,542,544,546
,548,550,552,554,556,558,560,562,564,566,568,570,572,574,576,578,580,582,584,585,586,588
,590,592,594,596,598,600,602,604,606,608,610,612,614,616,618,620,622,624,626,628,630,632
,635,637)] <- "factor"
z <- read.table("/Users/irenecho/Desktop/stat443/datafiles/subset.txt",header=TRUE,colCl
asses=vartype)
```

Question 1: IPW estimate using RPART

```
tmp <- rep(NA,nrow(z))
tmp[z$INTRDVX_ == "C"] <- 0
tmp[z$INTRDVX_ == "D" | z$INTRDVX_ == "T"] <- 1
z$INTRDVX_ <- tmp  ### convert INTRDVX to binary variable
rp <- rpart(INTRDVX_ ~ . - INTRDVX - FINLWT21, data=z, method="anova")
plot(rp,compress=TRUE,margin=0.1)
text(rp)
```

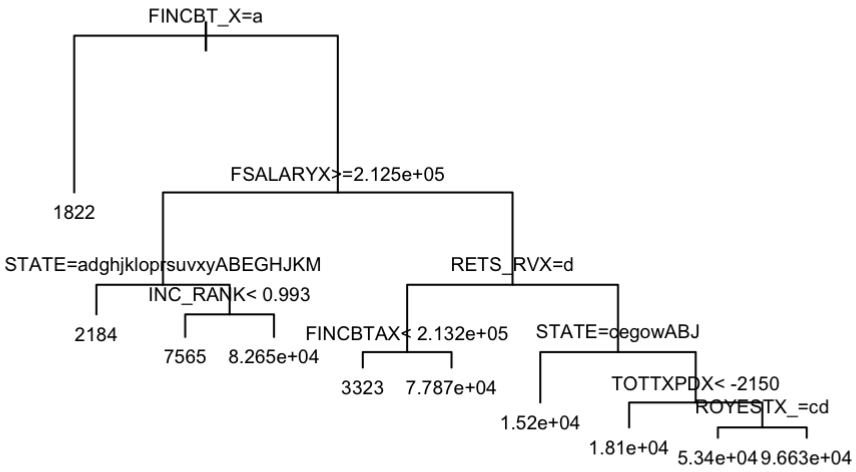



```
p <- predict(rp) ### predicted prob(INTRDVX_ = 1)
w <- z$FINLWT21
y <- z$INTRDVX
gp <- !is.na(y)
ipw <- sum(w[gp]*y[gp]/p[gp])/sum(w[gp]/p[gp])
print(ipw)
```

```
## [1] 4442.648
```

Question 2: Imputation estimate using RPART

```
rp2 <- rpart(INTRDVX ~ . - INTRDVX_, weight=FINLWT21, data=z,method="anova")
plot(rp2, compress = TRUE, margin=0.2)
text(rp2, cex=0.6)
```




```
y <- z$INTRDVX
w <- z$FINLWT21
miss <- is.na(y) ## obs with missing INTRDVX
yhat <- predict(rp2,newdata=z)
popmean <- (sum(w[!miss]*y[!miss])+sum(w[miss]*yhat[miss]))/sum(w)
print(popmean)
```

```
## [1] 3996.971
```

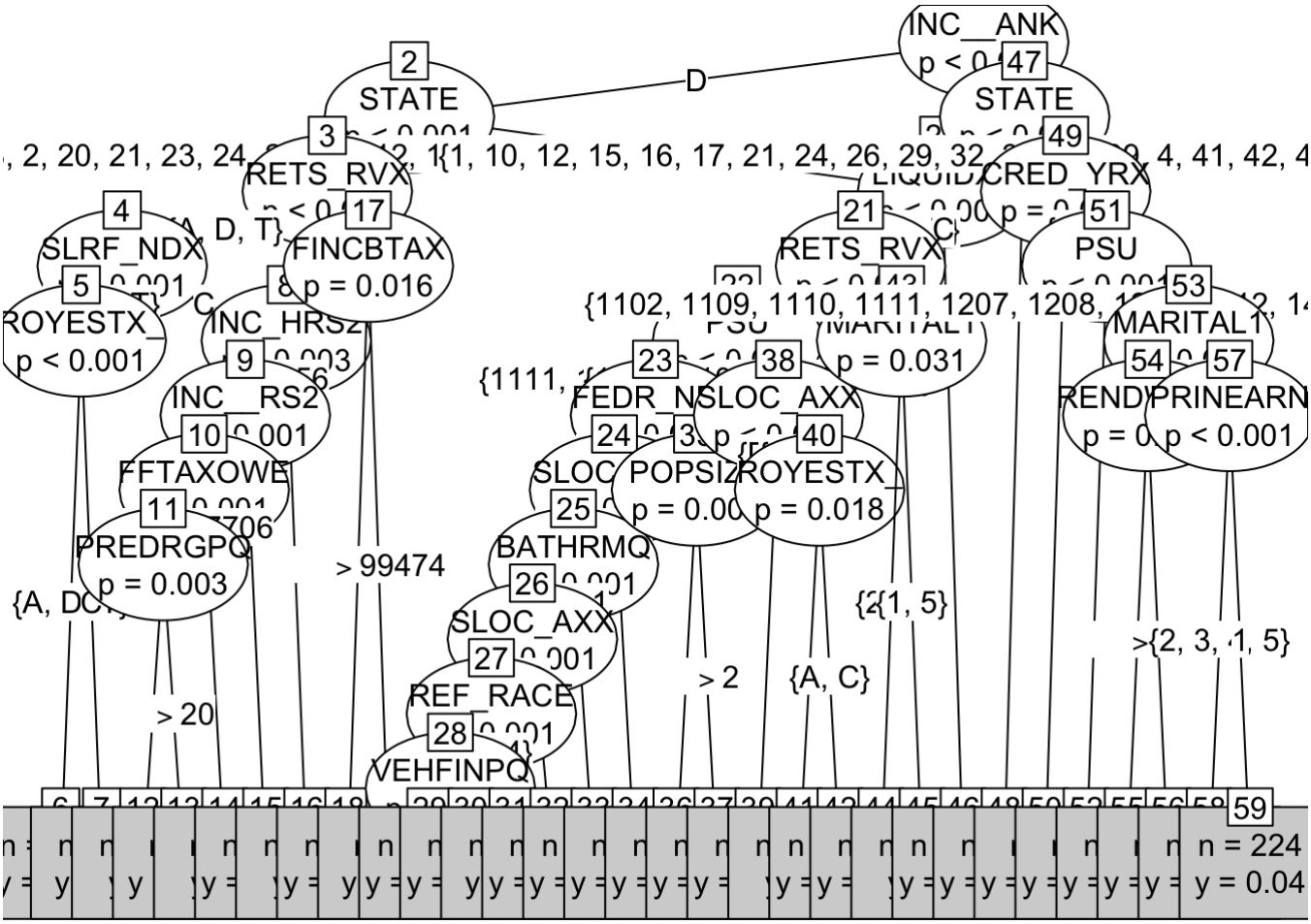
Question 3 - CTREE

```
vartype <- rep("numeric",638)
vartype[c(1,2,4,6,8,10,12,14,16,18,20,22,23,24,25,26,27,29,31,33,34,35,37,39,41,43,45,47
,49,52,54,56,58,60,62,64,66,68,70,72,73,74,75,76,77,78,79,80,82,84,86,88,89,90,92,94,96,
97,98,99,100,102,104,106,108,109,110,111,112,113,114,115,116,118,119,120,122,123,124,125
,126,128,130,131,132,133,135,137,139,303,304,305,306,307,308,309,310,311,312,313,314,315
,316,317,318,319,321,323,325,331,333,407,409,410,411,453,454,456,458,460,462,464,465,466
,467,468,470,472,474,476,477,478,479,482,484,486,488,490,492,494,496,497,498,499,500,502
,504,506,508,510,512,514,516,518,520,522,524,526,528,530,532,534,536,538,540,542,544,546
,548,550,552,554,556,558,560,562,564,566,568,570,572,574,576,578,580,582,584,585,586,588
,590,592,594,596,598,600,602,604,606,608,610,612,614,616,618,620,622,624,626,628,630,632
,635,637)] <- "factor"
z <- read.table("/Users/irenecho/Desktop/stat443/datafiles/subset.txt",header=TRUE,colCl
asses=vartype)
z <- z[-c(461)]

tmp <- rep(NA,nrow(z))
tmp[z$INTRDVX_ == "C"] <- 0
tmp[z$INTRDVX_ == "D" | z$INTRDVX_ == "T"] <- 1
z$INTRDVX_ <- tmp

### regression tree without INTRDVX and FINLWT21
fmla <- formula(INTRDVX_ ~ . - INTRDVX - FINLWT21, data=z)
ct <- party::ctree(fmla, data=z)
```

```
## Warning in factor_trafo(x): factors at only one level may lead to problems
## Warning in factor_trafo(x): factors at only one level may lead to problems
## Warning in factor_trafo(x): factors at only one level may lead to problems
## Warning in factor_trafo(x): factors at only one level may lead to problems
## Warning in factor_trafo(x): factors at only one level may lead to problems
## Warning in factor_trafo(x): factors at only one level may lead to problems
## Warning in factor_trafo(x): factors at only one level may lead to problems
## Warning in factor_trafo(x): factors at only one level may lead to problems
## Warning in factor_trafo(x): factors at only one level may lead to problems
## Warning in factor_trafo(x): factors at only one level may lead to problems
## Warning in factor_trafo(x): factors at only one level may lead to problems
## Warning in factor_trafo(x): factors at only one level may lead to problems
## Warning in factor_trafo(x): factors at only one level may lead to problems
## Warning in factor_trafo(x): factors at only one level may lead to problems
## Warning in factor_trafo(x): factors at only one level may lead to problems
## Warning in factor_trafo(x): factors at only one level may lead to problems
## Warning in factor_trafo(x): factors at only one level may lead to problems
## Warning in factor_trafo(x): factors at only one level may lead to problems
## Warning in factor_trafo(x): factors at only one level may lead to problems
## Warning in factor_trafo(x): factors at only one level may lead to problems
## Warning in factor_trafo(x): factors at only one level may lead to problems
## Warning in factor_trafo(x): factors at only one level may lead to problems
## Warning in factor_trafo(x): factors at only one level may lead to problems
## Warning in factor_trafo(x): factors at only one level may lead to problems
## Warning in factor_trafo(x): factors at only one level may lead to problems
## Warning in factor_trafo(x): factors at only one level may lead to problems
## Warning in factor_trafo(x): factors at only one level may lead to problems
```

```
## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems
```

```
plot(ct,type="simple",drop_terminal = TRUE)
```




```
y <- z$INTRDVX
p <- predict(ct)
gp <- !is.na(y)
ipw <- sum(w[gp]*y[gp]/p[gp])/sum(w[gp]/p[gp])
print(ipw)
```

```
## [1] 4445.513
```

CFOREST

```
fmla <- formula(INTRDVX_ ~ . - INTRDVX - FINLWT21)
cf <- party::cforest(fmla, data=z)
```

```
## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems
```

```
## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems
```

```
p <- predict(cf,newdata=z)
```

```
## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems
```

```
## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems

## Warning in factor_trafo(x): factors at only one level may lead to problems
```

```
w <- z$FINLWT21
y <- z$INTRDVX
gp <- !is.na(y)
ipw <- sum(w[gp]*y[gp]/p[gp])/sum(w[gp]/p[gp])
print(ipw)
```

```
## [1] 4689.638
```

Homework #4 – Kyungjin Cho
Extra Writing

Introduction:

There are three ways to estimate IPW: RPART, CTREE, and CFOREST.
Although we are getting the same IPW, the values differ a bit by which type of estimation we are using.

Difference between RPART, CTREE, and CFOREST

Three of them are R package which are used to model classification and regression trees. While in rpart we can only build one tree, the result is easy to interpret. However, in random forest, there are many trees that can be conducted and the result is produced by the combined effort of all the trees so it is not easily interpretable. Yet, since random forest implements ensemble learning algorithm, the accuracy is better than rpart package. Therefore, the predictive power of Random Forest is better than rpart.

Interpretation of the Result

In this homework, I have calculated IPW using RPART, CTREE, and CFOREST. Here are the estimated results of IPW:

	RPART	CTREE	CFOREST
IPW	4442.648	4445.513	4689.638

Conclusion

I found out that rpart was the easiest and fastest to compute among three estimators. Next was ctree and the last one was cforest. However, I realized that because of the ensemble learning algorithm, cforest is so far the best estimate.