STAT 443 HW 4

Caitlin Bolz

3/17/2021

```
library(rpart)
library(party)
```

Code for reading subset.txt into R

Data has been prepared/pre-processed in GUIDE. The descriptor for INTRDVX was changed from x to n. A data set was created that contains only the non-excluded (x) variables. Additionally the column with all NA's was removed.

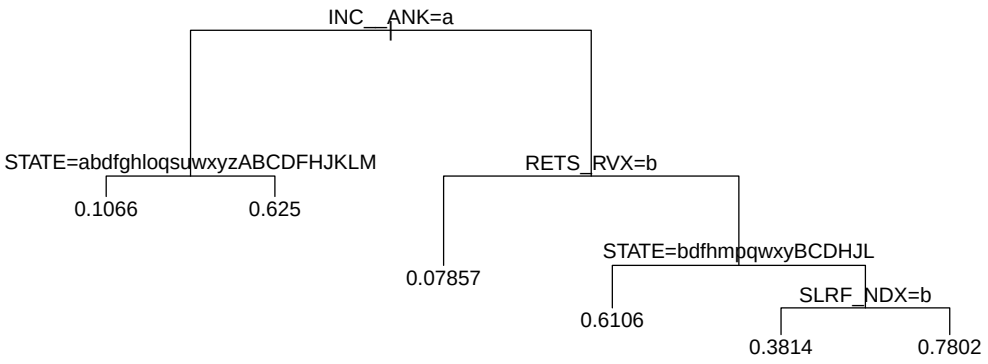
```
vartype = rep("numeric",638)
vartype[c(1,2,4,6,8,10,12,14,16,18,20,22,23,24,25,26,27,29,31,33,34,35,37,39,
          41,43,45,47,49,52,54,56,58,60,62,64,66,68,70,72,73,74,75,76,77,78,79,
          80,82,84,86,88,89,90,92,94,96,97,98,99,100,102,104,106,108,109,110,
          111,112,113,114,115,116,118,119,120,122,123,124,125,126,128,130,131,
          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("subset.txt",header=TRUE,colClasses=vartype)
z = z[, colSums(is.na(z)) < nrow(z)]
```

Code used in multiple parts

```
tmp = rep(NA,nrow(z))
tmp[z$INTRDVX_ == "C"] = 0
tmp[z$INTRDVX_ == "D" | z$INTRDVX_ == "T"] = 1
z$INTRDVX_ = tmp
y = z$INTRDVX
w = z$FINLWT21
```

Question 1

```
rp = rpart(INTRDVX_ ~ . - INTRDVX - FINLWT21, data=z, method="anova")
plot(rp,compress=TRUE,margin=0.1)
text(rp)
```

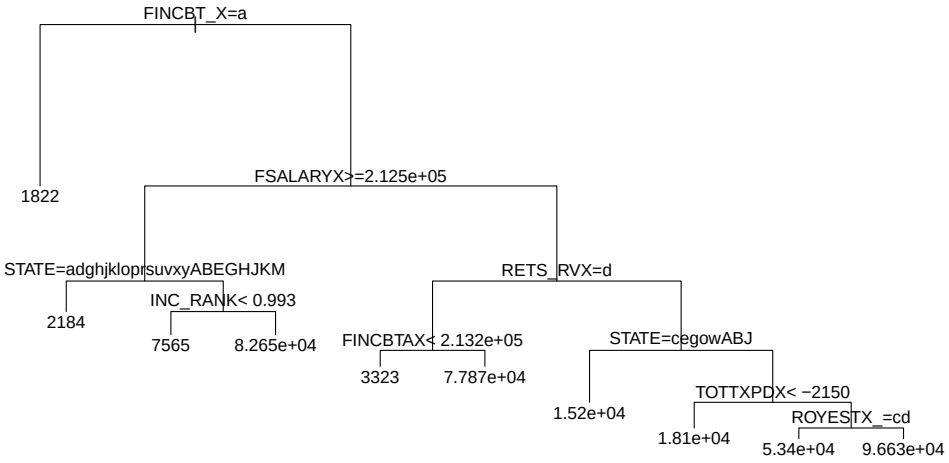



```
p = predict(rp)
gp = !is.na(y)
ipw = sum(w[gp]*y[gp]/p[gp])/sum(w[gp]/p[gp])
print(ipw)
```

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

Question 2

```
rp2 = rpart(INTRDVX ~ . - INTRDVX_, weight = FINLWT21, data = z, method = "anova")
plot(rp2, compress = T, margin = 0.1)
text(rp2)
```



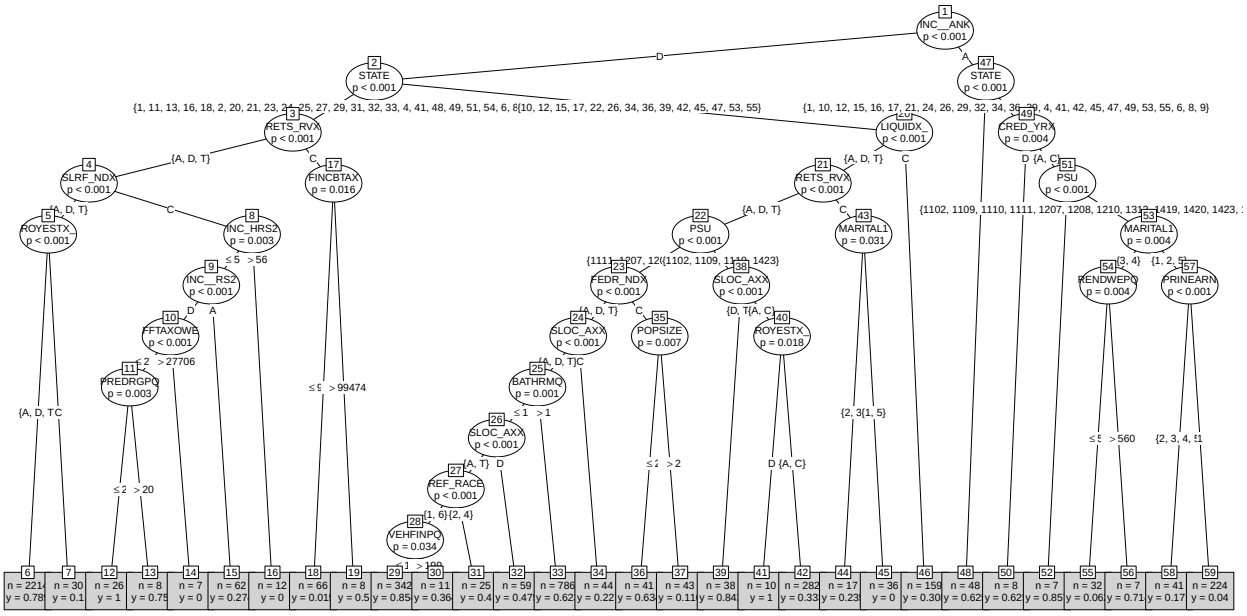

```
miss = is.na(y)
yhat = predict(rp2, newdata = z)
popmean = (sum(w[!miss]*y[!miss])+sum(w[miss]*yhat[miss]))/sum(w)
print(popmean)
```

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

Question 3

CTREE

```
fmla = formula(INTRDVX_ ~ . - INTRDVX - FINLWT21, data = z)
ct = ctree(formula = fmla, data=z)
plot(ct,type="simple",drop_terminal = TRUE)
```



```
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 = cforest(fmla, data=z)
p = predict(cf)
gp = !is.na(y)
ipw = sum(w[gp]*y[gp]/p[gp])/sum(w[gp]/p[gp])
print(ipw)
```

```
## [1] 4694.173
```