SURVMETH 745 HW9

Stacey Frank & Chendi Zhao

4/12/2022

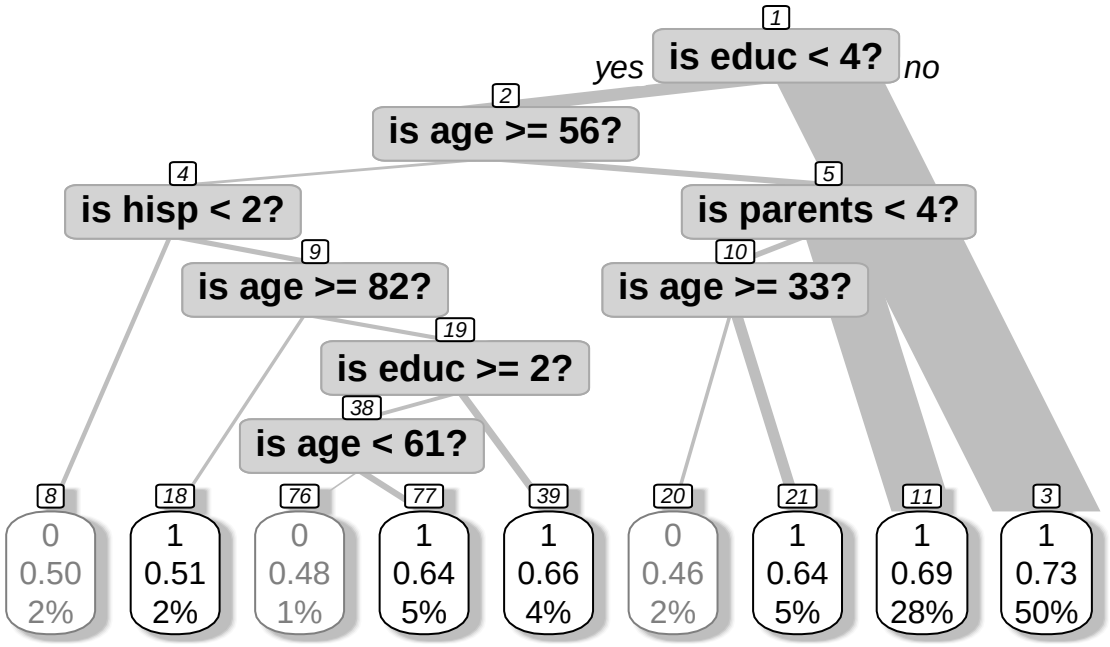
13.9

```
set.seed(15097)

cart1 <- rpart(resp ~ age + sex + hisp + race + parents + educ,
               method = "class",
               control = rpart.control(minbucket = 50, cp = 0),
               data = nhis)

print(cart1, digits = 4)

table(cart1$where)
```

This CART classification tree has nine terminal nodes. The first node is classified as all of those respondents and nonrespondents who have a GED or more education. The eight other nodes come from among the group that have less than a GED.The next split in the tree is between those who are aged 56 or older, and those younger than age 56. Of those who are younger than age 56, the tree divides those cases into three terminal nodes. Selected persons who do not have a parent in the family are put into one category. Of those who do have a parent present in the family, they are divided into two terminal nodes, one of people who are aged 33 or older, and those who are younger than 33.

Of those people who are aged 56 or older and have less than a GED education, those are divided into five terminal nodes. The first is those are Hispanic. Of those who are non-Hispanic, those who are aged 82 or older are divided into their own terminal node. For non-Hispanics who are between age 56 and 81, they are divided based on education again. Of those who have an education level of some high school education or more, they are divided on age, with cases who are aged 56-60 divided into one node and those who are aged 61-81 being divided into another node. The last node is those between the ages of 56 and 81 who have an education level of 8th grade or less. A visual representation of the classification tree is below.

```
## cex 1   xlim c(0, 1)   ylim c(0, 1)
```




```
#unweighted response rate by node

unwt.rr <- by (as.numeric(nhis[, "resp"]), cart1$where, mean)

unwt.rr
```

The unweighted response rates in each of the nine nodes that are formed are: 1) 50%, 2) 51%, 3) 48%, 4) 64%, 5) 66%, 6) 46%, 7) 64%, 8) 69%, and 9) 73%.

13.10

##a.Using classes identified from 13.9

```
#weighted response rate by node
wt.rr <- by(data = data.frame(resp = as.numeric(nhis[,"resp"]),
wt = nhis[,"svywt"]),
cart1$where,
function(x) {weighted.mean(x$resp, x$wt)} )


nhis.NR <- cbind(nhis, NR.class=cart1$where)
table(nhis.NR$NR.class, useNA="always")
```

```
##
```

```
##    4    6    9   10   11   14   15   16   17 <NA>
##   88   73   52  209  166   67  210 1099 1947    0
```

```
tmp1 <- cbind(NR.class=as.numeric(names(wt.rr)), unwt.rr, wt.rr)
nhis.NR <- merge(nhis.NR, data.frame(tmp1), by="NR.class")
nhis.NR <- nhis.NR[order(nhis.NR$ID),]

nhis.NR%>%
group_by(NR.class)%>%
summarise(alt3a=mean(unwt.rr),
alt4a=mean(wt.rr))
```

```
## # A tibble: 9 x 3
##   NR.class alt3a alt4a
##      <int> <dbl> <dbl>
## 1        4 0.5   0.519
## 2        6 0.507 0.502
## 3        9 0.481 0.480
## 4       10 0.636 0.642
## 5       11 0.663 0.679
## 6       14 0.463 0.453
## 7       15 0.643 0.645
## 8       16 0.692 0.703
## 9       17 0.731 0.747
```

##b. Using classes identified from alternative 1

```
p.class <- pclass(formula = resp~age +
as.factor(hisp) +
as.factor(race) +
as.factor(parents_r) +
as.factor(educ_r),
type = "unwtd", data = nhis, link="logit", numcl=5)
table(p.class$p.class, useNA="always")
```

```
##
## [0.453,0.631] (0.631,0.677] (0.677,0.714] (0.714,0.752] (0.752,0.818]
##           783           782           782           782           782
##          <NA>
##             0
```

##Alternative3 Unweighted response rate

```
alt3b=by(as.numeric(nhis[, "resp"]), p.class$p.class, mean)
```

##Alternative4 Weighted response rate

```
alt4b=by(data = data.frame(resp = as.numeric(nhis[,"resp"]),
wt = nhis[,"svywt"]),
 p.class$p.class,
function(x) {weighted.mean(x$resp, x$wt)})

cbind(alt3b,alt4b)
```

```
##                    alt3b     alt4b
## [0.453,0.631] 0.5862069 0.5897205
## (0.631,0.677] 0.6662404 0.6813254
## (0.677,0.714] 0.6930946 0.7007653
## (0.714,0.752] 0.7084399 0.7176318
## (0.752,0.818] 0.7966752 0.8041545
```

"Alt3a" and "Alt4a" were formed using nine classes identified from Exercise 13.9."Alt3a" and "Alt4a" were formed using five classes identified from Alternative1 in Example 13.8.

Both alternatives produced similar weights under the two scenarios. In practice, we need to consider the survey design when selecting the weights to use. If every unit in a class has the same probability of responding, i.e., the grouping is very effective, alternative 3 is preferred.

Alternative 4 is an estimate of the population response rate in class c assuming MAR. This estimate is approximately unbiased with respect to the compound sampling/response mechanism or with respect to a model with a common response probability within each class.This choice can be inefficient if weights vary much within class and units have a common propensity score for unit i.

Five classes are usually recommended based on some analyses in Cochran (1968). With a large sample, there is no reason not to create more classes. This can help make each more homogeneous on covariates and propensity scores. More classes may decrease bias due to nonresponse but may increase variances by creating bigger spread in the weights.