ACS Healthcare Data - Prep File

Intro to Data Science for Public Policy, Spring 2016

PPOL 670

Contents

For those of you who are interested in understanding how the healthcare data was prepped, review the code below.

Start by reading in the .zip file from the Census Bureau.

```
#Download
  url <- "https://www2.census.gov/programs-surveys/acs/data/pums/2015/1-Year/csv_pga.zip"
  temp <- tempfile()
  download.file(url, temp, mode="wb")
  unz <- unzip(temp)
  df <- read.csv(unz[1])
```

Upon downloading the data, we'll need to recode a few demographic varaibles. Healthcare coverage is the labeled variable (target).

```
#Keep people who are 16 or older
  df <- df[df$AGEP>=16,]

#Health coverage
  df$coverage <- NA
  df$coverage[df$HICOV == 2] <- "No Coverage"
  df$coverage[df$HICOV == 1] <- "Coverage"
  df$coverage <- as.factor(df$coverage)
```

Using the data dictionary, we'll restructure and label the education, citizenship, marriage, and race.

```
#Education
  df$educ[df$SCHL<16 ] <- "Less than HS"
  df$educ[df$SCHL>=16 & df$SCHL<21] <- "HS Degree"
  df$educ[df$SCHL==21] <- "Undergraduate Degree"
  df$educ[df$SCHL>21] <- "Graduate Degree"
  df$educ <- as.factor(df$educ)

#Citizenship
  df$cit[df$CIT != 5] <- "Citizen"
  df$cit[df$CIT == 5] <- "Non-citizen"
  df$cit <- as.factor(df$cit)

#Marriage
  df$mar[df$MAR == 1] <- "Married"
  df$mar[df$MAR == 2] <- "Widowed"
  df$mar[df$MAR == 3] <- "Divorced"
  df$mar[df$MAR == 4] <- "Separated"
  df$mar[df$MAR == 5] <- "Never Married"
  df$mar <- as.factor(df$mar)

#Race
```

```
  df$race[df$RAC1P == 1] <- "White"
  df$race[df$RAC1P == 2] <- "Black"
  df$race[df$RAC1P == 3] <- "Amer. Ind."
  df$race[df$RAC1P == 4] <- "Alaska Native"
  df$race[df$RAC1P == 5] <- "Tribes Spec."
  df$race[df$RAC1P == 6] <- "Asian"
  df$race[df$RAC1P == 7] <- "Nat. Hawaiian/Pac. Isl."
  df$race[df$RAC1P == 8] <- "Other"
  df$race[df$RAC1P == 9] <- "Two or More"
  df$race <- factor(df$race)
```

For consistency, we'll rename the `AGEP` and `WAGE` variables.

```
  colnames(df)[c(8,70)] <- c("age", "wage")
```

As only a fraction of respondents to the survey did not have healthcare coverage, we will need to "boost" the signal of this subpopulation in order to better emphasize differences. This usually is done by either (1) resampling the smaller subpopulation until there is a 1:1 ratio of the two subpopulations, or (2) undersampling the larger population. In this case, we will undersampled the people with coverage.

```
#Prep for undersampling 'covered'
  temp <- df[df$coverage=="Coverage",]

#Set seed to so that the random draw is replicable
  set.seed(123)
  health <- rbind( df[df$coverage=="No Coverage",],
                   temp[sample(row.names(temp), sum(df$coverage == "No Coverage")),])
```

Lastly, we only need to keep the following variables.

```
  vars <- c("coverage", "age", "wage", "cit", "mar", "educ", "race")
  health <- health[, vars]
  write.csv(health, "lecture7.csv", row.names = FALSE)
```