ACS Assignment 03/24

Yitong Wu

March 24, 2020

```
library(readr)
orig <- read.csv("org.csv")
syn1 <- read.csv("syn1.csv")
syn2 <- read.csv("syn2.csv")
syn3 <- read.csv("syn3.csv")
```

```
CalculateKeyQuantities <- function(origdata, syndata, known.vars, syn.vars, n){
  origdata <- origdata
  syndata <- syndata
  n <- n

  c_vector <- rep(NA, n)
  T_vector <- rep(NA, n)

  for (i in 1:n){
    match <- (eval(parse(text=paste("origdata$",syn.vars,"[i]==
                                    syndata$",syn.vars,sep="",collapse="&")))&
                eval(parse(text=paste("origdata$",known.vars,"[i]==
                                      syndata$",known.vars,sep="",collapse="&"))))
    match.prob <- ifelse(match, 1/sum(match), 0)

    if (max(match.prob) > 0){
      c_vector[i] <- length(match.prob[match.prob == max(match.prob)])
    }
    else
      c_vector[i] <- 0
      T_vector[i] <- is.element(i, rownames(origdata)[match.prob == max(match.prob)])
  }

  K_vector <- (c_vector * T_vector == 1)
  F_vector <- (c_vector * (1 - T_vector) == 1)
  s <- length(c_vector[c_vector == 1 & is.na(c_vector) == FALSE])

  res_r <- list(c_vector = c_vector,
                T_vector = T_vector,
                K_vector = K_vector,
                F_vector = F_vector,
                s = s
  )
  return(res_r)
}
```

```
IdentificationRisk <- function(c_vector, T_vector, K_vector, F_vector, s, N){

  nonzero_c_index <- which(c_vector > 0)

  exp_match_risk <- sum(1/c_vector[nonzero_c_index]*T_vector[nonzero_c_index])
```

```
  true_match_rate <- sum(na.omit(K_vector))/N
  false_match_rate <- sum(na.omit(F_vector))/s

  res_r <- list(exp_match_risk = exp_match_risk,
                true_match_rate = true_match_rate,
                false_match_rate = false_match_rate
  )
  return(res_r)
}
```

```
known.vars <- c("SEX", "RACE", "MAR")
syn.vars <- c("LANX", "WAOB", "DIS", "HICOV")
n <- dim(orig)[1]
```

```
name <- "syn"
csv <- ".csv"
exp_risk <- c()
true_rate <- c()
false_rate <- c()
for (i in 1:3){
  num <- toString(i)
  file <- paste(name,num,csv,sep="")
  syn <- read.csv(file)
  KeyQuantities <- CalculateKeyQuantities(orig, syn,
                                          known.vars, syn.vars, n)
  c_vector <- KeyQuantities[["c_vector"]]
  T_vector <- KeyQuantities[["T_vector"]]
  K_vector <- KeyQuantities[["K_vector"]]
  F_vector <- KeyQuantities[["F_vector"]]
  s <- KeyQuantities[["s"]]
  N <- n
  ThreeSummaries <- IdentificationRisk(c_vector, T_vector, K_vector, F_vector, s, N)
  exp_risk <- append(exp_risk, ThreeSummaries[["exp_match_risk"]])
  true_rate <- append(true_rate, ThreeSummaries[["true_match_rate"]])
  false_rate <- append(false_rate, ThreeSummaries[["false_match_rate"]])
}
```

```
mean(exp_risk)
```

```
## [1] 41.46743
```

```
mean(true_rate)
```

```
## [1] 0.0005666667
```

```
mean(false_rate)
```

```
## [1] 0.9638026
```