ACS Match Risks

Yitong Wu

March 3, 2020

```r
library(readr)
orig_dt <- read.csv("ACSdata_org.csv")
synt_dt <- read.csv("ACSdata_syn.csv")
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
orig_dt <- orig_dt[
  with(orig_dt, order(SEX, RACE, MAR)),
]
synt_dt <- synt_dt[
  with(synt_dt, order(SEX, RACE, MAR)),
]
```

```r
summary(orig_dt$SEX)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    1.00    2.00    1.53    2.00    2.00
```

```r
summary(orig_dt$RACE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.426   2.000   6.000
```

```r
summary(orig_dt$MAR)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   2.374   4.000   5.000
```

Expected Match Risk

```r
risk <- c()
for (i in 1:2){
  for (j in 1:6){
    for (k in 1:5){
      t_o <- orig_dt %>% filter(SEX==i, RACE==j, MAR==k)
      t_s <- synt_dt %>% filter(SEX==i, RACE==j, MAR==k)
      c <- dim(t_o)[1]
      t <- 0
      match <- t_o$LANX==t_s$LANX & t_o$WAOB==t_s$WAOB & t_o$DIS==t_s$DIS & t_o$HICOV==t_s$HICOV & t_o$
      t <- sum(match)
      }
```

```
      a <- t/c
      risk <- append(risk,a)
    }
  }
}
sum(risk)
```

```
## [1] 3.223269
```

True match risk

```
true <- c()
for (i in 1:2){
  for (j in 1:6){
    for (k in 1:5){
      t_o <- orig_dt %>% filter(SEX==i, RACE==j, MAR==k)
      t_s <- synt_dt %>% filter(SEX==i, RACE==j, MAR==k)
      c <- dim(t_o)[1]
      t <- 0
      k <- 0
      match <- t_o$LANX==t_s$LANX & t_o$WAOB==t_s$WAOB & t_o$DIS==t_s$DIS & t_o$HICOV==t_s$HICOV & t_o$
      t <- sum(match)
      }
      if (t*c==1){
        k <- 1
      }
      true <- append(true,k)
    }
  }
}
sum(true)
```

```
## [1] 0
```

Question: my c_i is always much greater than 1, then how can $t_i \times c_i$ be equal to 1?

False match risk

```
false <- c()
for (i in 1:2){
  for (j in 1:6){
    for (k in 1:5){
      t_o <- orig_dt %>% filter(SEX==i, RACE==j, MAR==k)
      t_s <- synt_dt %>% filter(SEX==i, RACE==j, MAR==k)
      c <- dim(t_o)[1]
      t <- 0
      f <- 0
      match <- t_o$LANX==t_s$LANX & t_o$WAOB==t_s$WAOB & t_o$DIS==t_s$DIS & t_o$HICOV==t_s$HICOV & t_o$
      t <- sum(match)
      }
      if ((1-t)*c==1){
        f <- 1
      }
      false <- append(false,f)
    }
  }
}
sum(false)
```

```
## [1] 0
```