ACS Data Identification Disclosure Risk

MATH 301 Data Confidentiality

Henrik Olsson

March 24, 2020

```
ACSdata_org <- read.csv("ACSdata_org.csv")
ACSdata_syn <- read.csv("ACSdata_syn.csv")
ACSdata_syn2 <- read.csv("ACSdata_syn2.csv")
ACSdata_syn3 <- read.csv("ACSdata_syn3.csv")
```

Note: The datasets exlucde the HISP variable in the ACS data dictionary, but every other variable stays the same with the same name and description.

The following four variables are synthesized: LANX, WAOB, DIS, HICOV.

Where m = 3 case

Step 1: calculate key quantities

- four synthesized variables: LANX, WAOB, DIS, HICOV, assigned to syn.vars
- three known variables: SEX, RACE, MAR, assigned to known.vars

```
known.vars <- c("SEX", "RACE", "MAR")
syn.vars <- c("LANX", "WAOB", "DIS", "HICOV")
n <- dim(ACSdata_org)[1]
KeyQuantities1 <- CalculateKeyQuantities(ACSdata_org, ACSdata_syn, known.vars, syn.vars, n)
KeyQuantities2 <- CalculateKeyQuantities(ACSdata_org, ACSdata_syn2, known.vars, syn.vars, n)
KeyQuantities3 <- CalculateKeyQuantities(ACSdata_org, ACSdata_syn3, known.vars, syn.vars, n)

## Step 2: calculate 3 summary measures

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)
}

## each record is a target, therefore N = n

c_vector1 <- KeyQuantities1[["c_vector"]]
T_vector1 <- KeyQuantities1[["T_vector"]]
K_vector1 <- KeyQuantities1[["K_vector"]]
F_vector1 <- KeyQuantities1[["F_vector"]]
s1 <- KeyQuantities1[["s"]]
N <- n
```

```
ThreeSummaries1 <- IdentificationRisk(c_vector1, T_vector1, K_vector1, F_vector1, s1, N)

c_vector2 <- KeyQuantities2[["c_vector"]]
T_vector2 <- KeyQuantities2[["T_vector"]]
K_vector2 <- KeyQuantities2[["K_vector"]]
F_vector2 <- KeyQuantities2[["F_vector"]]
s2 <- KeyQuantities2[["s"]]
N <- n
ThreeSummaries2 <- IdentificationRisk(c_vector2, T_vector2, K_vector2, F_vector2, s2, N)

c_vector3 <- KeyQuantities3[["c_vector"]]
T_vector3 <- KeyQuantities3[["T_vector"]]
K_vector3 <- KeyQuantities3[["K_vector"]]
F_vector3 <- KeyQuantities3[["F_vector"]]
s3 <- KeyQuantities3[["s"]]
N <- n
ThreeSummaries3 <- IdentificationRisk(c_vector3, T_vector3, K_vector3, F_vector3, s3, N)
```

Summaries:

```
## Expected match risk
(ThreeSummaries1[["exp_match_risk"]] + ThreeSummaries2[["exp_match_risk"]] + ThreeSummaries3[["exp_matc
```

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

```
## True match rate
(ThreeSummaries1[["true_match_rate"]] + ThreeSummaries2[["true_match_rate"]] + ThreeSummaries3[["true_ma
```

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

```
## False match rate
(ThreeSummaries3[["false_match_rate"]] + ThreeSummaries3[["false_match_rate"]] + ThreeSummaries3[["fals
```

```
## [1] 0.9640288
```

Results and Discussion

```
41.46743/10000
```

```
## [1] 0.004146743
```

The expected match risk is 0.00414 probability average for each record to be correctly identified

```
0.0005666667*10000
```

```
## [1] 5.666667
```

5.666667, or approximately 6 records are correct unique matches

```
(s1 * ThreeSummaries3[["false_match_rate"]] + s2 * ThreeSummaries3[["false_match_rate"]] + s3 * ThreeSu
```

```
## [1] 155.2086
```

The 0.964 false match rate: among the 195 (the value of each s1, s2, s3) unique matches, 155 are false matches.