Disclosure Risk

Reese Guo

2/17/2020

Propensity Score Measure

```
load("Rdata.RData")
df <- data.frame(Income = logIncome_syn, syn = 1, expend = logExpenditure)
df1 <- data.frame(Income = logIncome, syn = 0, expend = logExpenditure)
Data <- rbind(df, df1)
N <- nrow(Data_rand)
c <- 1/2

logistic <- glm(syn ~ Income + expend, data = Data, family = "binomial")
summary(logistic)
```

```
##
## Call:
## glm(formula = syn ~ Income + expend, family = "binomial", data = Data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.298  -1.180   0.007   1.176   1.329
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.72144    0.53754  -1.342   0.1796
## Income       0.10128    0.04010   2.526   0.0115 *
## expend      -0.04078    0.05407  -0.754   0.4508
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2756.0  on 1987  degrees of freedom
## Residual deviance: 2749.5  on 1985  degrees of freedom
## AIC: 2755.5
##
## Number of Fisher Scoring iterations: 3
```

```
slope_income <- 0.10128
slope_expend <- -0.04078
intercept <- -0.72144

income <- Data[,1]
expenditure <- Data[,3]
d <- income * slope_income + + expenditure * slope_expend + intercept
p <- 1 / (1 + d)
diff <- (p-c)^2
U_p <- sum(diff) / N
U_p
```

```
## [1] 0.2803906
```

Cluster Analysis

```
Data_rand <- Data[sample(nrow(Data)),]
size <- 10
G <- ceiling(N/size)
n_o <- c()
for(i in 0:G){
  new_data <- Data_rand[(i * 10 + 1):((i + 1) * 10),]
  syn <- new_data$syn
  n_o <- c(n_o, sum(syn == 0))
}
n_o <- n_o[1:198]
diff_cluster <- (n_o/size - c)^2
U_c <- sum(diff_cluster) / G
U_c
```

```
## [1] 0.0258794
```

Empirical CDF

```
S_x <- ecdf(logIncome)
S_y <- ecdf(logIncome_syn)
S_diff <- c()
for(i in 1:length(logIncome)){
  S_diff <- c(S_diff, (logIncome[i] - logIncome_syn[i])^2)
}

U_m <- sqrt(max(S_diff))
U_m
```

```
## [1] 6.41631
```

```
U_s <- sum(S_diff) / N
U_s
```

```
## [1] 1.376762
```

Disclosure Risk

For this part, I will be using the New York City Airbnb data. Non-Id or non-geographic information in this dataset include name of the host, listing space type, price in dollars, amount of nights minimum, number of reviews, date of latest view, number of review per month, amount of listing per host, and number of days when listing is available for booking. Geographic information include neighborhood by group, neighborhood by area, and longitude and latitude information.

Of all the information included in the dataset, I think number of days when listing is avaiable for booking is a sensitive information because it suggests what a property is used for. I divided the number of days available for booking to four categories: 331 days to 360 days, which suggests that the property is solely a rental property, 271-330 days, which suggests that the property may be used as vacation house and is listed

for rental in time not in use, 61-270 days, which suggests a property is sometimes used by the owner and sometimes used for rental, and 0-60 days, which suggests that a property is mainly used by its owner.

Since most of the other information are available on Airbnb website or application, we will assume that a potential intruder has information about a property's neighborhood by group and area, room type, and number of reviews. With the above-mentioned assumptions established, we will investigate the indentification risk of an intruder correctly identifying the category in days when listing is available for booking. For the simplicity of this assignment, we will use neighborhood by group and room type.

```
bnbData <- read.csv("AB_NYC_2019.csv")
length <- dim(bnbData)
avail <- bnbData$availability_365
Category <- c()

for(i in 1:length){
  if (avail[i] > 330){
    Category <- c(Category, 1)
  }else if( avail[i] <= 330 & avail[i] > 270){
    Category <- c(Category, 2)
  }else if( avail[i] <= 270 & avail[i] > 60){
    Category <- c(Category, 3)
  }else{
    Category <- c(Category, 4)
  }
}
```

```
## Warning in 1:length: numerical expression has 2 elements: only the first
## used
```

```
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
```

```
library("data.table")
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
bnbData_cat <- data.frame(bnbData, category = Category)
neigh_unique <- unique(bnbData_cat$neighbourhood_group)
room_unique <- unique(bnbData_cat$room_type)

Data_select <- bnbData_cat[, c("neighbourhood_group", "room_type", "category")]
filter(Data_select, neighbourhood_group == "Manhattan", room_type == "Private room") %>% group_by(categ
```

```
## # A tibble: 4 x 2
```

```
##   category count
##      <dbl> <int>
## 1        1   847
## 2        2   568
## 3        3  2049
## 4        4  4518
```

```
N_0 <- nrow(bnbData_cat)

p_1 <- 847/N_0
p_1
```

```
## [1] 0.01732283
```

```
p_2 <- 568/N_0
p_2
```

```
## [1] 0.01161673
```

```
p_3 <- 2049/N_0
p_3
```

```
## [1] 0.04190613
```

```
p_4 <- 4518/N_0
p_4
```

```
## [1] 0.09240209
```

The above numbers are the risks that a property in Manhattan that is a private room will disclose its days of available listing information in this dataset. If we assume that the intruder knows more information, these disclose risk will increase.