Prevalence of Having Contraband among Pulled-over Drives

Yukun Li

12/4/2019

Background and Introduction

To be done.

Causal Roadmap

Scientific Question:

What is the prevalence of having contraband if all drives are searched.

Causal Model

W_1 : age

W_2 : race

W_3 : gender

W_4 : vehicle type

Δ : if search is conducted

Y* : Underlying contraband status

Y : if contraband is found

- Endogenous variables: $X = (W_1, W_2, W_3, W_4, \Delta, Y)$
- Exogenous variables: $U \sim \mathbb{P}_U$ (to be determined).

Structral equation F:

$$
\begin{aligned}
W_1 &= f_{W1}(U_{W1}) \\
W_2 &= f_{W2}(U_{W2}) \\
W_3 &= f_{W3}(U_{W3}) \\
W_4 &= f_{W4}(W_1, W_2, W_3, U_{W4}) \\
\Delta &= f_{\Delta}(W_1, W_2, W_3, W_4, U_{\Delta}) \\
Y^* &= f_{Y^*}(W_1, W_2, W_3, W_4, U_{Y^*}) \\
Y &= \Delta \times Y^*
\end{aligned}
$$

Causal Parameter

$\Psi^*(\mathbb{P}^*) = \mathbb{P}^*(Y^* = 1) = \mathbb{P}^*(Y_{\Delta=1})$

Observed data and its link to causal model

Observed data are randomly generated from the structual causal model.

Identifiability

Lest's assume all Us are independent.

Statistical estimand

$\Psi^*(\mathbb{P}^*) = \Psi(\mathbb{P}_0) = \mathbb{E}_W\{\mathbb{P}_0(Y=1|\Delta=1,W)\}$

Estimate

Parametric G-computation (simple substitution estimator), IPTW, TMLE. Use super learner during the estimating procedure. Don't forget to talk about the positivity assuptions.

Present a detailed plan for statistical inference/variance estimation based on the non-parametric bootstrap and implement it.

Data preprocessing

```r
# packages
library(tidyverse)
library(lubridate)
```

```r
# load dataset
dat <- readRDS("data/MAStatePatrol.rds")
# take a look at the variables we have
colnames(dat)
```

```
##  [1] "raw_row_number"             "date"
##  [3] "location"                   "county_name"
##  [5] "subject_age"                "subject_race"
##  [7] "subject_sex"                "type"
##  [9] "arrest_made"                "citation_issued"
## [11] "warning_issued"             "outcome"
## [13] "contraband_found"           "contraband_drugs"
## [15] "contraband_weapons"         "contraband_alcohol"
## [17] "contraband_other"           "frisk_performed"
## [19] "search_conducted"           "search_basis"
## [21] "reason_for_stop"            "vehicle_type"
## [23] "vehicle_registration_state" "raw_Race"
```

```r
# the dataset is balanced over years,
# we will use the observations only in 2015 for
# computational convenience and intepretability of results.
table(year(dat$date))
```

```
##
##   2007   2008   2009   2010   2011   2012   2013   2014   2015
## 247357 468131 428714 388280 335974 418846 400931 384468 343537
```

```r
dat_prep <- function(dat, loc, years){

  datBos <-  dat %>%
  filter(location == loc) %>%
  filter(year(date) == years) %>%
  filter(subject_race != 'unknown' & subject_race != 'other') %>% # positivity assumption
  filter(vehicle_type != 'Motorcycle' & vehicle_type != 'Trailer') %>% # positicity assumption
```

```
  filter(!is.na(subject_age) & !is.na(subject_sex)) %>%
  select(subject_age,
         subject_race,
         subject_sex,
         vehicle_type,
         contraband_found,
         search_conducted,
         # the following variables are not used.
         outcome,
         frisk_performed,
         search_basis,
         reason_for_stop,
         raw_Race)
# Here I select all the varibales that might be useful.
# A further discussion is needed to decide how to use them.

# drop unused levels from the dataframe
datBos <- droplevels(datBos)
return(datBos)
}

datBos <- dat_prep(dat, 'BOSTON', 2014)

# show summary and check postivity assumptions
summary(datBos$search_conducted)
```

```
##    Mode   FALSE    TRUE
## logical   41139     253
```

```
summary(datBos$contraband_found)
```

```
##    Mode   FALSE    TRUE    NA's
## logical     156      97   41139
```

```
summary(datBos$subject_age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.00   28.00   37.00   39.13   49.00   94.00
```

```
summary(datBos$subject_sex)
```

```
##   male female
##  30358  11034
```

```
table(datBos$subject_race)
```

```
##
## asian/pacific islander                  black               hispanic
##                   4004                   8153                   4441
##                  white
##                  24794
```

```
table(datBos$vehicle_type)
```

```
##
##   Commercial    Passenger Taxi/Livery
##         2581        36610        2201
```

```r
###############!!!!!###############
# positivity assumptions are heavily violated if we include vehicle type in W.
table(datBos$subject_race, datBos$subject_sex, datBos$vehicle_type, datBos$search_conducted)
```

```
## , ,  = Commercial,  = FALSE
##
##
##                           male female
##   asian/pacific islander   131      3
##   black                    240     10
##   hispanic                 337     23
##   white                   1754     79
##
## , ,  = Passenger,  = FALSE
##
##
##                           male female
##   asian/pacific islander  2586    771
##   black                   4910   2091
##   hispanic                2933    988
##   white                  15141   6943
##
## , ,  = Taxi/Livery,  = FALSE
##
##
##                           male female
##   asian/pacific islander   496      5
##   black                    814     18
##   hispanic                 101     11
##   white                    705     49
##
## , ,  = Commercial,  = TRUE
##
##
##                           male female
##   asian/pacific islander     0      0
##   black                      0      0
##   hispanic                   0      0
##   white                      4      0
##
## , ,  = Passenger,  = TRUE
##
##
##                           male female
##   asian/pacific islander    11      1
##   black                     60     10
##   hispanic                  38     10
##   white                     95     22
##
## , ,  = Taxi/Livery,  = TRUE
##
##
##                           male female
##   asian/pacific islander     0      0
```

```
##   black                       0      0
##   hispanic                    0      0
##   white                       2      0
```

```
table(datBos$subject_race, datBos$subject_sex, datBos$search_conducted)
```

```
## , ,  = FALSE
##
##
##                           male female
##   asian/pacific islander  3213    779
##   black                   5964   2119
##   hispanic                3371   1022
##   white                  17600   7071
##
## , ,  = TRUE
##
##
##                           male female
##   asian/pacific islander    11      1
##   black                     60     10
##   hispanic                  38     10
##   white                    101     22
```

Following are some attemps of estiamtion, I'll consider more parametric models and super learner and put all of them in R functions.

G-computation

```
# G-computation
# (1) NPMLE
# (2) logistic model : E(Y|D, W) ~ all Ws
# (3) may inlcude some interaction.
datBos.searched <-  datBos %>% filter(search_conducted == TRUE)
fit.gcomp <- glm(contraband_found ~
                   subject_age +
                   as.factor(subject_race) +
                   subject_sex +
                   as.factor(vehicle_type),
                 family = 'binomial', data = datBos.searched)
datBos.intervene <- datBos %>% mutate(search_conducted = TRUE)
EY.gcomp <- predict(fit.gcomp, newdata = datBos.intervene, type = 'response')
est.gcomp <- mean(EY.gcomp)
est.gcomp
```

```
## [1] 0.3506859
```

IPTW

```
# P0(A = 1|W) = logit^(-1){B0 + B1W1 + B2W2 + B2W3 + B4W4}
fit.prob.D <- glm(search_conducted ~
                    subject_age +
                    as.factor(subject_race) +
```

```
                    subject_sex +
                    as.factor(vehicle_type),
                  family = 'binomial', data = datBos)
prob.D1 <- predict(fit.prob.D, type = 'response')
summary(prob.D1)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0001001 0.0026199 0.0048569 0.0061123 0.0086241 0.0270126
```

```
# calculate weights
wt1 <- as.numeric(datBos$search_conducted == 1)/prob.D1
summary(wt1)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.000    0.000    0.000    1.026    0.000 1130.208
```

```
est.IPTW <- mean(wt1*datBos$contraband_found, na.rm = TRUE)
est.IPTW # too large
```

```
## [1] 55.0261
```

```
# Stabelized IPTW
wt.mean <- mean(wt1[!is.na(datBos$contraband_found)])
est.sIPTW <- mean(wt1*datBos$contraband_found, na.rm = TRUE)/wt.mean
```

TMLE