Intro, Search Rate, Metric, Conclusion

2022-07-06

Contents

Introduction

Traffic stops are one of the most common and routine ways in which Americans interact with the police; each year more than 20 million Americans are stopped for traffic violations [@daviswhydelangton2018]. Not all traffic stops are conducted equally, however. There is evidence of racial discrimination in traffic stops, as Black drivers are stopped more often than White driers on average [@pierson2020]. As such, federal and state mandates have increasingly required the collection of traffic stop data [@russell2001racial].

The Stanford Open Policing Project (SOPP), started by the Stanford Computational Journalism Lab and the Stanford Computational Policy Lab, has collected and standardized over 200 million records of traffic stop and search data from across the United States starting from 2015. In total, the SOPP contains 88 datasets consisting of data from both municipal police and state patrol agencies. The SOPP has been subject to analysis from myriad researchers who utilize the extensive range of traffic data for their research. However, much of the literature that references the SOPP does not consider how missing data might affect their analysis. In many publications, researchers have dealt with missing data simply by ignoring it. Our project aims to characterize missing data, specifically as it relates to race, to see how it could potentially influence analysis of the datasets. As we explore how to characterize missingness in the datasets, we pose three questions that guide our analysis: Are there fundamental differences between the traffic stop observations with high and low missingness? Does missingness have a certain trend with respect to race? Most importantly, are the missingness trends drastic enough to render two subsets of the same dataset incomparable?

Our project uses traffic-stop data from 23 city and state-wide datasets in the SOPP. Each dataset contains data on several variables that were collected during traffic stops such as driver age, driver race, stop location, etc. However, it is important to note that not every dataset contains the same variables and even if the variables are the same, they are often recorded at different rates (i.e. subject_race is recorded 99.9% of the time in Arizona but 87.1% of the time in Colorado). Our project focuses primarily on two variables: subject_race and search_conducted and explores patterns in missing data, specifically when there are missing values for the subject_race variable. Our aim for this project is to analyze search patterns in the missing data to inform a potential range of combinations for race among various racial groups. This new range of data could supplement existing indicators of racial bias such as unequal search rates for various racial groups (calculated without missing data), but it also has the potential to make us reconsider how missing data could alter tests performed on the dataset that include the subject_race variable.

What we used

In total, the SOPP contains 88 datasets, but we only use datasets that contain subject_race and search_conducted, narrowing down our list to only 23 datasets. The follwing list includes the location of the dataset as well as what we named the dataset in parentheses: Arizona (AZ), San Diego (CAsd), San Francisco (CAsf), Stockton (CAstockton), Colorado (CO), Connecticut (CT), Illinois (IL), Louisville (KYlouisville), New Orleans (LAno), Massachusetts (MA), Maryland (MD), Saint Paul (MNsaintpaul), Missouri (MO), Montana (MT), North Carolina (NC), Ohio (OH), Philadelphia, (PAph), Rhode Island (RI), South Carolina (SC), Nashville (TNna), San Antonio (TXsa), Vermont (VT), Wisconsin (WI).

Folder containing city/state datasets: https://drive.google.com/drive/folders/1IJUvwlzzymzQPDaEZaKCWhbdGIG4nfGF

How to import data:

```
# Make a vector of all your file paths
file_paths <- list.files(path = ("traffic_data/"), pattern = ".rds", full.names = TRUE)

# Make a vector of file names
file_names <-  gsub(pattern = "\\.rds$", replacement = "", x = basename(file_paths))

# Read all your data into a list
data_list <- lapply(file_paths, readRDS)

# To access individual datasets, index into the data list (i.e. data_list[[1]] accesses the first datas
```

Search rate

One way that racial profiling can be seen in traffic stops is through search rates that vary by race. When looking at city and state-wide datasets across the United States, it is apparent that search rates between White and Black populations vary substantially (it is important to note that our project focuses primarily on comparing White and Black populations for ease of calculation and proof of principle, but similar analysis can be applied to compare White and other minority populations). Our process was to first find the search rate of a few individual states and then calculate a search rate that encompassed all of the datasets in our data list.

We first removed all of the rows where search_conducted was not recorded (wherever 'NA' appeared in the search_conducted column) because we didn't know the search outcome in these cases. Then, we found the proportion of the entire population that was searched (dataset search rate). We used this value as a point of reference for the Black and White search rates, which we found next. We noticed that the White search rate was higher than the dataset search rate; whereas, the Black search rate was lower than the dataset search rate. Further, when directly comparing the Black and White search rates, we noticed that the Black search rate was over twice as high as the White search rate. The following code is an example of how we analyzed search proportions in a specific dataset, which is Colorado in this case:

```
CO <- readRDS("cities:states/CO.rds")

# remove rows where search_conducted was not recorded
CO_no_search_missing <- CO[!is.na(CO$search_conducted),]

# proportion of the entire population that was searched (0.35%)
CO_search_prop <- CO_no_search_missing %>%
  group_by(search_conducted) %>%
  summarize(searched = n()) %>%
  mutate(prop = searched/nrow(CO_no_search_missing))
```

```r
# proportion of white ppl who were searched (0.31%) -> under state search rate
CO_search_prop_white <- CO_no_search_missing %>%
  group_by(subject_race, search_conducted) %>%
  filter(subject_race == "white") %>%
  summarize(white_searched = n()) %>%
  mutate(prop = white_searched/nrow(CO_no_search_missing %>% filter(subject_race == "white")))

# proportion of black ppl who were searched (0.78%) -> above state search rate
CO_search_prop_black <- CO_no_search_missing %>%
  group_by(subject_race, search_conducted) %>%
  filter(subject_race == "black") %>%
  summarize(black_searched = n()) %>%
  mutate(prop = black_searched/nrow(CO_no_search_missing %>% filter(subject_race == "black")))
```

When looking at other areas in the country, we notice a similar pattern: not only is the Black search rateusually above the dataset search rate and the White search rate below, but the black search rate was nearly twice as high as the white search rate in the datasets that we examined. We explore this pattern by making a plot of White vs Black search rates. Each point on the plot includes the White search rate (as the x coordinate) and the Black search rate (as the y coordinate). In the end, we include over 20 relevant points, which comprised of datasets that had sufficient data on both subject_race and search_conducted. After plotting all of the points, we add a least squares regression line to the plot and find the resulting slope to be 1.67. This value reaffirms what we saw in Colorado: there is evidence of racial bias in search rates. Black individuals are being searched at a rate that is 1.67 times higher than white individuals.

```r
get_sc_rate <- function(state) {

  size <- nrow(state)

  missing = ifelse(sum(is.na(state %>% select(subject_race))) == 0, FALSE, TRUE)

  ifelse(size > 1000000,
         temp <- state %>% slice_sample(n = 1000000) %>% filter(!is.na(search_conducted)),
         temp <- state %>% filter(!is.na(search_conducted)))


  temp %>% group_by(subject_race, search_conducted) %>%
    filter(subject_race == "white" | subject_race == "black") %>%
    filter(search_conducted == TRUE) %>%
    summarise(searched = n()) %>%
    mutate(prop = case_when(subject_race == "black" ~ searched / nrow(temp %>% filter(subject_race == "
                            subject_race == "white" ~ searched / nrow(temp %>% filter(subject_race == "
    select(-search_conducted, -searched) %>%
    pivot_wider(names_from = subject_race, values_from = prop) %>%
    mutate(size = size) %>%
    mutate(missing = missing)
}


sc_rates <- map_dfr(data_list, get_sc_rate) %>% bind_cols(name = file_names)

sc_rates <- sc_rates %>% filter(name != "MN")

sc_plot <- sc_rates %>%
```

```
  ggplot(aes(x = white, y = black)) +
  geom_point(aes(size = `size`, color = `missing`, label = name)) +
  # ggrepel::geom_label_repel(aes(label = name)) +
  geom_smooth(method = "lm", se = FALSE) +
  ggtitle("Black vs White search rates") +
  labs(caption = "search_rate_black = 0.016 + 1.67 * search_rate_white")

sc_rates %>%
  lm(black ~ white, data = .) %>%
  tidy()

sc_plot

ggplotly(sc_plot)
```

Missingness Metric

We use odds ratios as the main metric to analyze missingness in our project. Odds ratios are calculated exactly like how one would expect; they are simply a ratio of two odds. Odds themselves are similar to probabilities except that the denominator is not the number of instances; rather, the denominator is the number of failures while the numerator is the number of successes.

Odds

In our project, we use odds in the context of the search_conducted variable. We calculate odds by dividing the total number of individuals who were NOT searched (the successes) by the total number of individuals who WERE searched (the failures).

```
# $\text{odds} = \text{searched} / \text{not searched}$
```

Odds Ratio (comparing race)

Our primary analysis utilizes odds ratios comparing Black and White populations through the lens of the search_conducted variable. We calculate this odds ratio by dividing the odds of not being searched for Black individuals by the odds of not being searched for White individuals. For example, if the odds ratio equals 1.3, the odds of not being searched as a Black individual is 1.3 times higher than for a White individual.

Odds Ratio (race vs NA)

To incorporate missingness into our analysis, we use odds ratios comparing the odds of not being searched in a population that we do know against the odds of not being searched in a population that we don't know. For example, if we want to calculate an odds ratio analyzing missing data as it relates to the Black population, we would set the numerator as the odds of not being searched in the Black population and the denominator as the odds of not being searched for those who didn't have their race recorded.

Below is the code for a function that finds two odds ratios for a particular dataset: odds_black / odds_NA and odds_white / odds_NA

```r
get_OR <- function(state) {

  # points are grouped by size in plot
  size <- nrow(state)

  temp <- state %>%
    select(subject_race, search_conducted) %>%
    filter(subject_race == "black" | subject_race == "white" | is.na(subject_race)) %>%
    mutate(search_conducted = as.logical(search_conducted)) %>% # convert to TRUE/FALSE
    group_by(search_conducted, subject_race) %>%
    summarize(Count = n()) %>%
    spread(key = search_conducted, value = Count, fill = 0) %>%
    rename(no_search = `FALSE`,
         search = `TRUE`) %>%
    mutate(odds = no_search / search) %>%
    select(subject_race, odds)
  # checks to see if there is missingness
  if (sum(is.na(state %>% select(subject_race))) != 0) {
    # pulls odds for NA values
    odds_NA <- temp %>%
      filter(is.na(subject_race)) %>%
      pull(odds)

    if (!is.infinite(odds_NA )) {

      temp <- temp %>%
        mutate(OR = odds/odds_NA) %>%
        select(subject_race, OR) %>%
        # make columns be levels from subject_race and values be odds ratios
        pivot_wider(names_from = subject_race, values_from = `OR`) %>%
        mutate(size = size) %>%
        select(-`NA`)

      return(temp)
    }
  }

  temp <- as_tibble_row(c(black = 0, white = 0)) %>%
    mutate(size = size)
}
```

The map function applies the get_OR function to each of the datasets.

```r
# map function returns a data frame consisting of odds ratios for each dataset created by row-binding (
or_sc <- map_dfr(data_list, get_OR) %>% bind_cols(name = file_names)

# includes all odds ratios except those where odds_black = 0 or odds_white = 0
or_sc <- or_sc %>% filter(!(black == 0 & white == 0))
```

Interpreting odds ratios vs NA plot

Assuming that odds ratio comparing white population to missing population is on the x-axis:

The y = x line tells us how the black and white search rates differ or if they are the same. If the point is located under the line, we know that the white search rate is higher than the black search rate; if the point lies above the line, we know that the black search rate is higher than the white search rate. If the point lies on the y = x line, we know that the black and white search rates are equal.

If a point lies on the x = 1 line, the odds of not being searched for those who didn't have their race recorded is the same as the odds of not being searched for white individuals. This indicates that search_conducted data for those who didn't have their race recorded is more similar to data for white individuals than black individuals (unless the point intersects the y = 1 line as well).

Similarly, if a point lies on the y = 1 line, the odds of not being searched for those who didn't have their race recorded is the same as the odds of not being searched for black individuals, suggesting that the search data for those who didn't have their race recorded is more similar to data for black individuals.

Code for plotting all of the places from data list onto an odds ratio plot:

```
or_plot <- or_sc %>%
  filter(name != "MT") %>%
  ggplot(aes(white, black)) +
  geom_point(aes(size = size, label = name)) +
  # y = x line
  geom_abline(slope = 1, intercept = 0, alpha = 0.25, linetype = "dashed") +
  # y = 1 line
  geom_hline(yintercept = 1, alpha = 0.25) +
  # x = 1 line
  geom_vline(xintercept = 1, alpha = 0.25) +
  scale_shape(solid = FALSE) +
  ggtitle("Search Conducted", subtitle = "Odds Ratios") +
  # adds labels to points
  ggrepel::geom_text_repel(aes(label = name)) +
  ylim(0,3.2) +
  xlim(0,3.2)

or_plot


ggsave(plot = or_plot, file = "or_plot.png", width = 14)

# point size depends on population size
plotly::ggplotly(or_plot)
```

As we can see on the plot, all of the points lie below the y = x line, meaning that the black search rate is higher than the white search rate in every place included in the data list. Further, some points lie close to the x = 1 line and y = 1 line which indicate that their missing data looks similar to the search data for either white or black individuals.

Log Odds

Another way that odds can be used in the context of one's analysis is by calculating log odds in a logistic regression. Taking the log of an odds ratio returns the general equation of the logistic regression.

Because we will never know the exact values of the missing data, we are never quite certain of the exact values which are to be inputted into the odds ratio comparing White populations to Black populations. This uncertainty, in turn, has the potential to alter the logistic regression. In the following analysis, we will explore potential ways to narrow down the values of this odds ratio (meaning how to narrow down the

total Black and White populations separated by search and accounting for the missing data), yet the sheer number of possible values (which depends on the amount of missingness) has the potential to render the logistic regression completely different altogether.

Amount of missingness -> extent of variation of NA rates -> how much logistic regression could be impacted

Conclusion

As we think back to the guiding questions posed in the introduction, we can see the extent to which missingness plays a role in analyzing datasets.

1. Are there fundamental differences between the traffic stop observations with high and low missingness?

Yes, this is demonstrated through the min/max dotplot and the NA rate boxplots. Datasets with more missingness like CO and WI have larger OR ranges on the dotplot compared to states like AZ and CAsd which have less missingness. Further, the extent to which the NA rates vary depends on the amount of missingness (more variation caused by more missingness in general).

2. Does missingness have a certain trend with respect to race?

With our current knowledge, we can't quite answer this question yet. We have looked at boxplots of NA rates in a few individual datsets, but we haven't looked at all datasets to gauge whether or not there is a missingness trend with respect to race.

3. Most importantly, are the missingness trends drastic enough to render two subsets of the same dataset incomparable?

Yes, we can show this by using the dotplot representing ORs of the datasets as well as the boxplots representing NA rates (bs, bns, ws, wns) for individual states. The possible OR can be anywhere in betweeen the min and max on the dotplot. If one were to compare the min and max OR and use the resulting NA rates, they would have the same total missingness but the subsequent analysis of the dataset could be drastically impacted becaue the NA rates would be different. Thus, without knowing the exact combination of NA rates, resulting analyses of the datset can vary substantially. Further, the extent to which the analysis can vary also depends on the amount of missingness in the datset (more missingness means more possible combinations of NA rates).

Ways to continue project

- find a way to transform the min/max dotplot into a collection of boxplots (could divide it up into parts to make it more easily viewable like city vs state patrol and then CO could be its own bc it skews the rest)
- connect odds_NA plot with later analysis (this might be possible by adding points to the min, max dotplot/boxplot that represent the x=1 and y=1 lines that would show that the missing data looked very similar to either the White or the Black population after missingness)
- answer guiding question #2: see if there is a trend throughout all of the datasets (like in general are the IQR and mean of the boxplots for bs and bns higher than for ws and wns <- this could show that there is probably more black missingness than white missingness proportion-wise in the datasets)
- find a way to connect age with race and search (and could potentially use rsa simulation)?
- apply work to other minority groups and see how analysis changes