Implications of Missing Traffic Stop Data

Amber Lee advised by Jo Hardin
Pomona College

November 5, 2021

Motivation

Concern over racial profiling has led to mass collection of traffic stop data. As the data has been made available to the public, researchers have modeled the data to look for evidence of discrimination.

Motivation

Concern over racial profiling has led to mass collection of traffic stop data. As the data has been made available to the public, researchers have modeled the data to look for evidence of discrimination.

- Most statistical models do not handle missing values, so most studies simply use complete-case analysis (`na.rm = True`).
- Chanin and Welsh (2020) draw attention to the issue of (non-random) missing values.

Motivation

Concern over racial profiling has led to mass collection of traffic stop data. As the data has been made available to the public, researchers have modeled the data to look for evidence of discrimination.

- Most statistical models do not handle missing values, so most studies simply use complete-case analysis (`na.rm = True`).
- Chanin and Welsh (2020) draw attention to the issue of (non-random) missing values.

$\Rightarrow$ **What are the trends of missing values in traffic stop data?**

Row-wise missingness

Let $X = (x_{ij})$ be a **dataset** with n observations and k variables.

- Each observation $i = 1, 2, \ldots, n$ represents a single traffic stop.
- We denote a **missingness indicator function** as

$$\mathbb{1}_M(x_{ij}) = \begin{cases} 1 & \text{if } x_{ij} \text{ is missing} \\ 0 & \text{if } x_{ij} \text{ is observed} \end{cases}$$

Row-wise missingness

Let $X = (x_{ij})$ be a **dataset** with n observations and k variables.

- Each observation $i = 1, 2, \ldots, n$ represents a single traffic stop.
- We denote a **missingness indicator function** as

$$\mathbb{1}_M(x_{ij}) = \begin{cases} 1 & \text{if } x_{ij} \text{ is missing} \\ 0 & \text{if } x_{ij} \text{ is observed} \end{cases}$$

Definition

The **stop missingness rate for observation** i (or row-wise SMR) is the percentage of missing values for a single traffic stop i.

$$\mathrm{SMR}_i = \frac{1}{k} \sum_{j=1}^{k} \mathbb{1}_M(x_{ij})$$

Dataset missingness

Pomona College

Definition

The **stop missingness rate for dataset** X (or dataset SMR) is the percentage of missing values in the entire dataset. Equivalently, it is the average row-wise missingness.

$$\mathrm{SMR}(X) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{SMR}_i$$

Dataset missingness

Definition

The **stop missingness rate for dataset** X (or dataset SMR) is the percentage of missing values in the entire dataset. Equivalently, it is the average row-wise missingness.

$$\mathrm{SMR}(X) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{SMR}_i$$

Example

$$\mathrm{SMR}(X) = 2/16$$

subject_race	subject_sex	subject_age	search_conducted
black	male	NA	0
other	male	22	0
hispanic	male	NA	0
black	male	35	0

Missingness by a variable

Definition

The SMR **for observation** i **restricted by variable** j' is the percentage of missing values for a traffic stop i, excluding variable j'.

Let $j' \in \{1, 2, \ldots, k\}$ be the column index for a variable of interest. The SMR for observation i restricted by j' is given by

$$\mathrm{SMR}_{i,j'} = \frac{1}{k-1} \sum_{j \neq j'} \mathbb{1}_M(x_{ij}).$$

Missingness by a variable

Definition

The SMR **for observation** i **restricted by variable** j' is the percentage of missing values for a traffic stop i, excluding variable j'.

Let $j' \in \{1, 2, \ldots, k\}$ be the column index for a variable of interest. The SMR for observation i restricted by j' is given by

$$\mathrm{SMR}_{i,j'} = \frac{1}{k-1} \sum_{j \neq j'} \mathbb{1}_M(x_{ij}).$$

Example

$$\mathrm{SMR}_{1,\texttt{race}} = 1/3$$
$$\mathrm{SMR}_{2,\texttt{race}} = 0$$

subject_race	subject_sex	subject_age	search_conducted
black	male	NA	0
other	male	22	0
hispanic	male	NA	0
black	male	35	0

SMR by `race`

Let's quantify how missingness varies by `race`.

Let $\texttt{r} \in \{1, 2, \ldots, k\}$ be the column index corresponding to the `race` variable in X.

Assume that $(x_{i,\texttt{r}})$ has only three levels: "White", "Other", and `NA` (missing). We can partition the observations into index sets (W), (O), and $(\texttt{NA}) \subseteq \{1, 2, \ldots, n\}$ such that

$$x_{i,\texttt{r}} = \begin{cases} \text{"White"} & i \in (W) \\ \text{"Other"} & i \in (O) \\ \texttt{NA} & i \in (\texttt{NA}). \end{cases}$$

SMR by race

The White-SMR, Other-SMR, and NA-SMR are given by:

$$\mathrm{SMR}_{(W)} = \frac{1}{|(W)|} \sum_{i \in (W)} \mathrm{SMR}_{i,\mathtt{r}}$$

$$\mathrm{SMR}_{(O)} = \frac{1}{|(O)|} \sum_{i \in (O)} \mathrm{SMR}_{i,\mathtt{r}}$$

$$\mathrm{SMR}_{(\mathtt{NA})} = \frac{1}{|(\mathtt{NA})|} \sum_{i \in (\mathtt{NA})} \mathrm{SMR}_{i,\mathtt{r}}.$$

Example

$$\mathrm{SMR}_{(B)} = 1/6$$

$$\mathrm{SMR}_{(H)} = 1/3$$

$$\mathrm{SMR}_{(O)} = 0$$

subject_race	subject_sex	subject_age	search_conducted
black	male	NA	0
other	male	22	0
hispanic	male	NA	0
black	male	35	0

SMR by `date` and `time`

We apply a similar method to continuous variables.

We need to be careful with partitioning the indices – the partitions need enough observations for the average restricted SMRs to be meaningful.

SMR by `date` and `time`

We apply a similar method to continuous variables.

We need to be careful with partitioning the indices – the partitions need enough observations for the average restricted SMRs to be meaningful.

- For `date`, we partition observations by the week.
- For `time`, we partition observations by the month and day/night.

About the datasets

The Stanford Open Policing Project has 100 million traffic stop observations from 56 police departments (so, 56 datasets) from 1999 to 2020. Each datasets records a different set of variables.

About the datasets

The Stanford Open Policing Project has 100 million traffic stop observations from 56 police departments (so, 56 datasets) from 1999 to 2020. Each datasets records a different set of variables.

Data pre-processing

We select $k = 9$ variables about

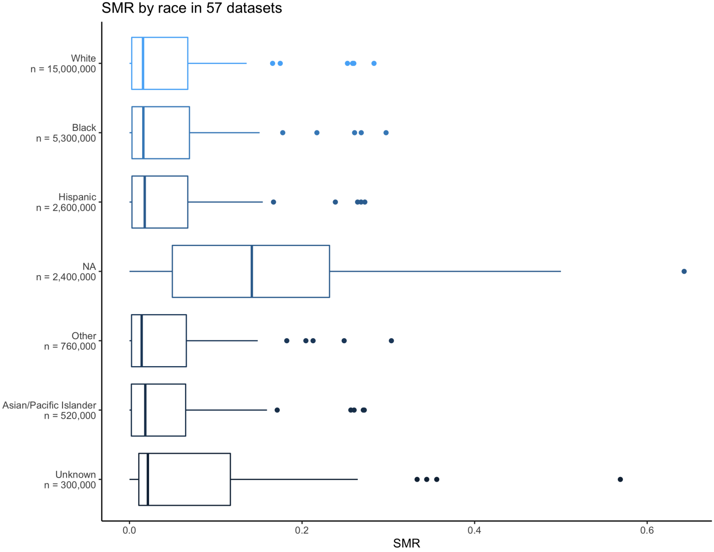
- driver demographic: `race`, `sex`, `age`;
- situational details: `time`, `date`, `latitude`, `longitude`; and
- outcomes: `search_conducted` and `arrest_made`.

And, for computational reasons, I take a 30% random sample of each dataset.

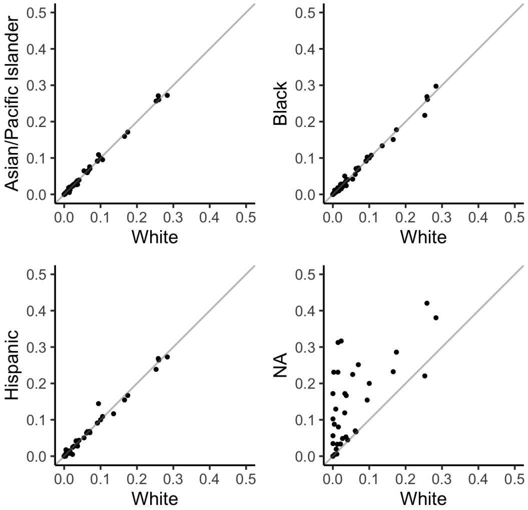
Section 3

Visualizations

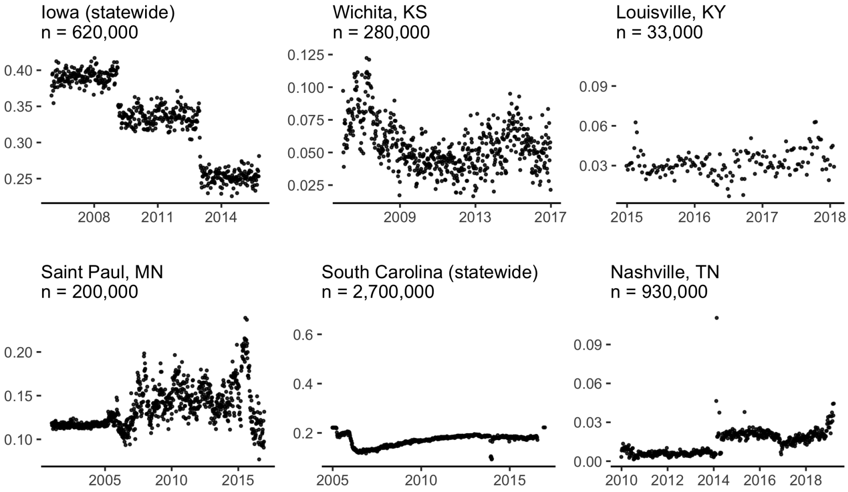
SMR by race


SMR by race in 57 datasets
White
n = 15,000,000
Black
n = 5,300,000
Hispanic
n = 2,600,000
NA
n = 2,400,000
Other
n = 760,000
Asian/Pacific Islander
n = 520,000
Unknown
n = 300,000
0.0
0.2
0.4
0.6
SMR

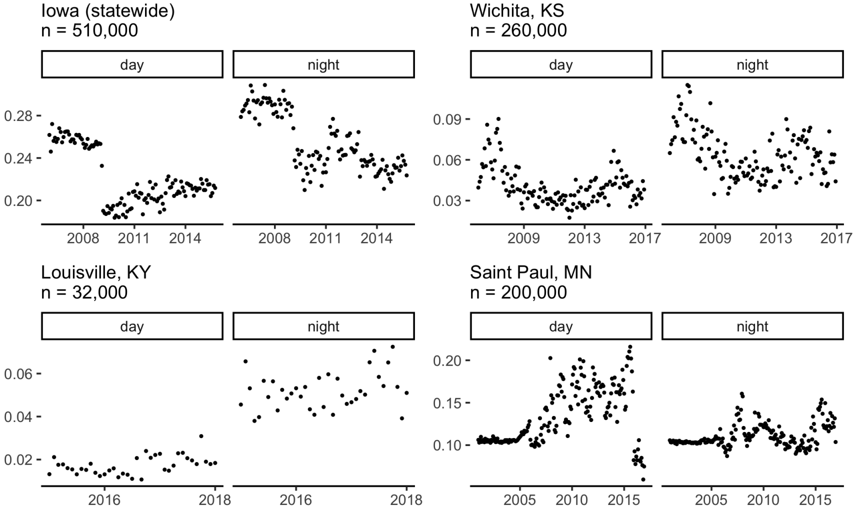
SMR by race


Asian/Pacific Islander
Black
Hispanic
NA
White
0.0
0.1
0.2
0.3
0.4
0.5

SMR by date


Iowa (statewide)
n = 620,000
0.40
0.35
0.30
0.25
2008
2011
2014
Wichita, KS
n = 280,000
0.125
0.100
0.075
0.050
0.025
2009
2013
2017
Louisville, KY
n = 33,000
0.09
0.06
0.03
2015
2016
2017
2018
Saint Paul, MN
n = 200,000
0.20
0.15
0.10
2005
2010
2015
South Carolina (statewide)
n = 2,700,000
0.6
0.4
0.2
2005
2010
2015
Nashville, TN
n = 930,000
0.09
0.06
0.03
0.00
2010
2012
2014
2016
2018

SMR by time


Iowa (statewide)
n = 510,000
day
night
0.28
0.24
0.20
2008
2011
2014
Wichita, KS
n = 260,000
0.09
0.06
0.03
2009
2013
2017
Louisville, KY
n = 32,000
0.06
0.04
0.02
2016
2018
Saint Paul, MN
n = 200,000
0.20
0.15
0.10
2005
2010
2015

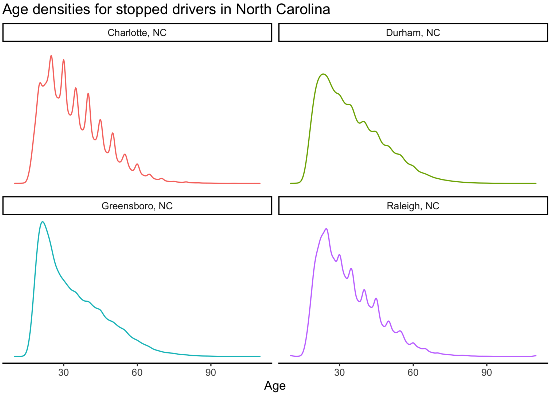
Discussion

- Regarding `race`, there isn't much evidence that Asian/Pacific Islander, Black, and Hispanic traffic stops are recorded differently than White traffic stops.
- Traffic stops with missing `race` have higher rates of missing values in other variables, too.
- Traffic stops are recorded differently based on the `date` and `time` of the traffic stop. Exactly *how* the missingness is different depends on the dataset.
- Future research could investigate the missingness mechanism (missing at random and missing not at random).

Thank you!

Note: Data collection is a deeply human process.

In states like California and New York, officers use *perception* to record driver race, gender, and age.



Evidently, North Carolina uses perception for age, too.