

Implications of Missing Traffic Stop Data

Amber Lee advised by Jo Hardin
Pomona College

November 5, 2021



Table of Contents



1 Definitions

2 About the data

3 Visualizations

Motivation



Public concern over racial profiling has led to the collection of traffic stop data. As traffic stop data has been made available to the public, researchers have tried to quantitatively look for evidence of discrimination.

Motivation



Public concern over racial profiling has led to the collection of traffic stop data. As traffic stop data has been made available to the public, researchers have tried to quantitatively look for evidence of discrimination.

- A generic model may look like $\text{search_conducted} \sim \text{race} + \text{gender} + \text{age} + \text{day/night}$.
- 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



Public concern over racial profiling has led to the collection of traffic stop data. As traffic stop data has been made available to the public, researchers have tried to quantitatively look for evidence of discrimination.

- A generic model may look like `search_conducted ~ race + gender + age + day/night`.
- 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.

⇒ **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 $i = 1, 2, \dots, n$ 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 $i = 1, 2, \dots, n$ 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 .

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

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.

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

Example

$$\text{SMR}(X) = 7/12$$

Missingness by a variable

One limitation of SMR_i and $\text{SMR}(X)$ is that we can't see how missingness varies *by* a variable, like time of day or the race of the driver.

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

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

Example

SMR by race



Let's quantify how missingness varies by race.

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

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

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

SMR by race



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

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

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

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

Example

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

- amasses 100 million number of traffic stop observations
- from 21 state patrol agencies and 35 municipal police departments
- from 1999 to 2020.

Data pre-processing



We consider 32 total datasets and $k = 9$ variables.

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

Note: data collection is a deeply human process.



In states like California and New York, officers use *perception*(!!!) to gauge driver race, gender, and age. Evidently, North Carolina definitely uses perception for age, too.

Missingness by race



yeehaw

Post-stop outcomes by race



yeehaw

Missingness by date



graphic

Missingness by time



graphic

Discussion



- race
- post-stop outcomes by race, NA race drivers
- date
- day vs. night