SURP 2022: Characterizing Missing Traffic Stop Data

Saatvik Kher PO ‘24; Kyle Torres PO ‘25; Mentor: Dr. Jo Hardin

Introduction

Our project utilizes data from Stanford Open Policing Project (SOPP) datasets which contain traffic stop data from both municipal police and state patrol agencies in the United States. Our goal for this project is to analyze search patterns in missing traffic 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.

Logistic Regression & OR

Much of the literature uses SOPP data to train logistic regressions without considering how the missing data might affect their analysis. We use logistic regressions to define odds ratios (OR) and examine the impact of missingness on OR.

$$odds_{black} = \frac{\#\ not\ searched}{\#\ searched} \Big\} \ black\ population$$

$$odds_{white} = \frac{\#\ not\ searched}{\#\ searched} \Big\} \ white\ population$$

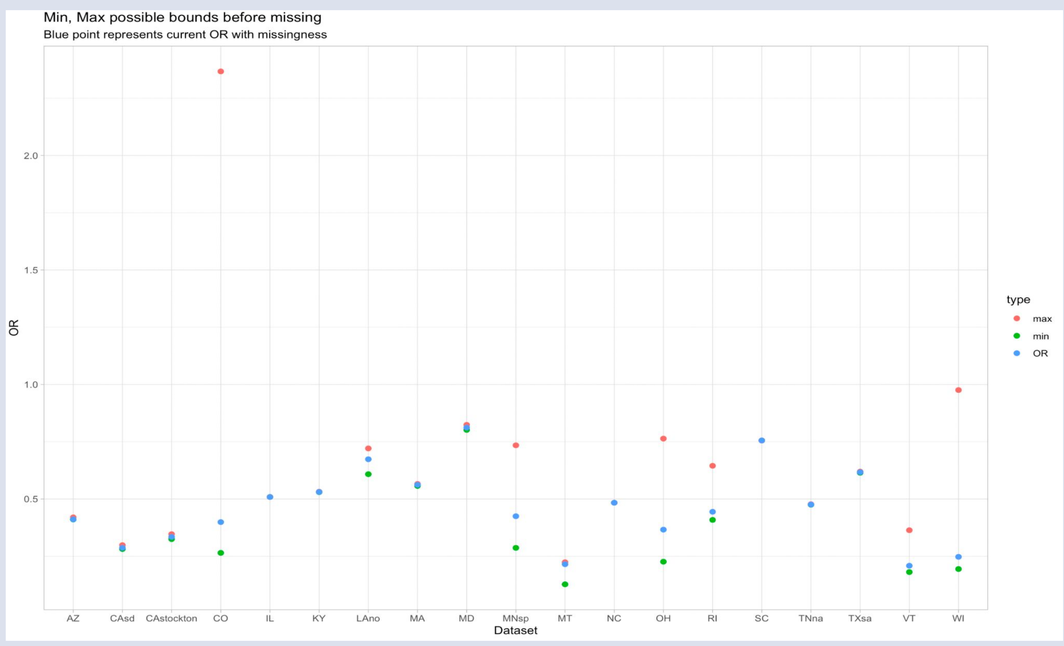
$$odds_{NA} = \frac{\#\ not\ searched}{\#\ searched} \Big\} \ missing\ race$$

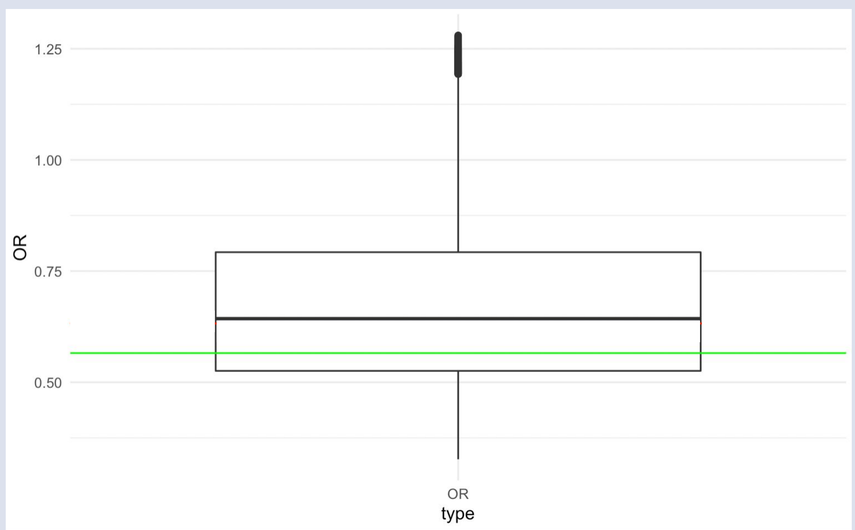
Hence we define:

$$OR_{black} = \frac{odds_{black}}{odds_{NA}}$$

$$OR_{white} = \frac{odds_{white}}{odds_{NA}}$$

We assess the impact of different missingness structures on OR to see if missingness trends can be drastic enough to render two subsets of the same dataset incomparable?





“True” OR

The figures represent a range of possible true ORs (with “true” implying that we knew the exact data that was missing) for the states and cities we considered. CO and WI display concerning trends because the range of possible true ORs is so large. The boxplot examines a simulated population where the truth is known and shows that the true OR is different than the OR after missingness.

References

Baumgartner, F. R., Epp, D. A., Shoub, K., & Love, B. (2017)
Goel, S., Rao, J. M., & Shroff, R. (2016)
Howell, D. C. (2007)
Little, R. J. A. (1988)
Manna, S., & Bunyard, S. (2021)
Pierson, E et al. (2020)
Riddell, C et al. (2020)
Rivera, R., & Rosenbaum, J. (2020)

Conclusions and Future Directions

1. There are fundamental differences between traffic stop observations with high and low missingness.
2. Patterns of missing racial data in datasets are drastic enough to potentially render two subsets of the same dataset incomparable.

Future directions include analyzing data for racial groups beyond white and black populations and further exploring the extent of missing racial data for each race to inform racial bias

Funded by: Posse Foundation Scholarship Fund & The Summer Undergraduate Research Program (SURP)

Saatvik Kher: suka2020@mymail.pomona.edu
Kyle Torres: kdta2021@mymail.pomona.edu