

STA234 HW7

Derin Gezgin

2025-04-29

Data Problem [30 points]

D.a

State your research questions and justify why are these research questions are valid using existing research in the field? You can update these based on the work you have done so far.

My current research questions are:

Q1: What is the relationship between being stopped and the general demographics?

Q2: How does the outcome of the traffic stop (warning, citation, search, arrest) relate to the racial demographics of the driver? Are the racial demographics significant predictors?

Q3: How does the amount of drivers stopped vary by time of the day and day of the year?

Q4: Are certain parts of SF have higher traffic-stop rates?

Pierson et al. (2020) and Xu et al. (2024) also has a similar approach to this research area where they focus on possible racial biases in traffic stops including the effect of time of the day the stop occurred [Q1-3].


In my study, I employ a similar approach where I visualize and test if there is any racial bias in the traffic stops. Different than Pierson et al. (2020), I focus on San Francisco, rather than using all the data from the USA.

It is important to note that, in my literature review so far, I could not find any studies that worked on the locational aspect of the traffic stops [Q4]. At the same time, most of the studies focused on the racial bias, while I am also planning to examine other demographical factors and if there is any statistically significant relationship between them and any of the possible response variables.

D.b

Perform the analysis you shared in hw-6 as possible ways to analyze your data to find answers for your research questions. What conclusions can you draw from this analysis?

Data Preparation

In this part, I do my default data preparation steps in order to prepare my data for the further analysis. 

```
# Reading the Data
traffic.data = read.csv("ca_sf_vehicle_2007_2016.csv")
race.data = read.csv("population_race_data.csv")

# Preparing the Variables
traffic.data$subject_sex = factor(traffic.data$subject_sex,
                                  levels = c("male", "female"),
                                  labels = c("Male", "Female"))

traffic.data$subject_race = factor(traffic.data$subject_race,
                                   levels = c("asian/pacific islander",
                                               "black",
                                               "hispanic",
                                               "white",
                                               "other"),
                                   labels = c("Asian/Pacific Islander",
                                              "Black",
                                              "Hispanic",
                                              "White",
                                              "Other"))

traffic.data$date = as.Date(traffic.data$date)
traffic.data$time = strptime(traffic.data$time, format = "%H:%M:%S")

traffic.data$outcome[is.na(traffic.data$outcome)] = "No Action"

traffic.data$outcome = factor(traffic.data$outcome,
                              levels = c("citation",
                                          "warning",
                                          "arrest",
                                          "No Action"),
                              labels = c("Citation",
                                         "Warning",
                                         "Arrest",
                                         "No Action"))

traffic.data$Year = format(traffic.data$date, "%Y")
```

```
# Merging the traffic data and the race data
traffic.data = merge(traffic.data,
                     race.data,
                     by = c("subject_race", "Year"),
                     all.x = TRUE)

colnames(traffic.data)[ncol(traffic.data)] = "proportioned_value"

traffic.data$Year = NULL

# Normalizing the proportions
traffic.data$generalWeights = (1 / (traffic.data$proportioned_value)) /
sum(1/(traffic.data$proportioned_value))
```

Outcome (Hit Rate) Test

I initially learned about this test from Pierson et al. (2020). In this paper, the researchers applied this to multiple counties and looked at the general picture across the USA, while as my research project focuses on the San Francisco data, I will apply the hit rate test only to my dataset.

I can start by filtering the traffic stops that resulted in a search and followed by further filtering this dataset to see the actual hit rate.

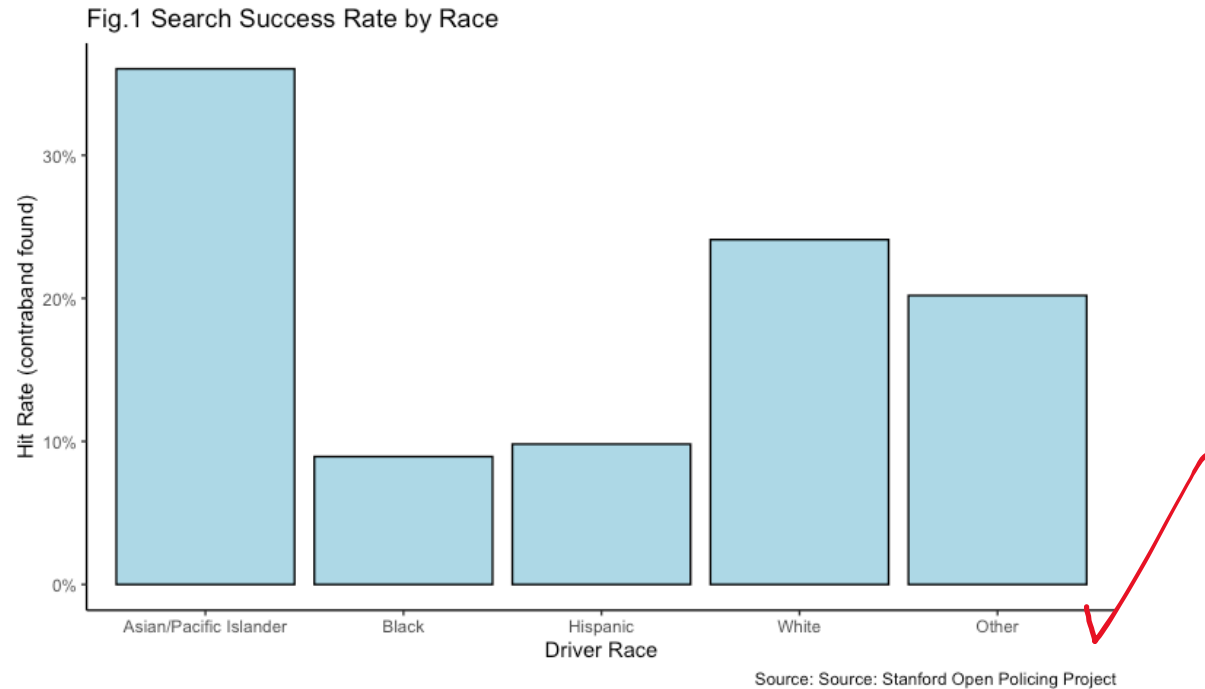
```
searches = traffic.data %>%
  filter(search_conducted == 1) %>%
  mutate(subject_race = factor(subject_race),
         found_contraband = as.integer(found_contraband))

hit_rates = searches %>%
  group_by(subject_race) %>%
  summarise(n_searches = n(),
            n_hits = sum(found_contraband),
            hit_rate = n_hits / n_searches)
```

After the preparation of the data, I can plot it to show the hit rate for different race groups

```
hitRatePlot = ggplot(data = hit_rates,
                    aes(x = subject_race,
                       y = hit_rate)) +
  geom_col(fill = "lightblue",
           color = "black") +
  scale_y_continuous(labels = scales::percent) +
  labs(x = "Driver Race",
       y = "Hit Rate (contraband found)",
       title = "Fig.1 Search Success Rate by Race",
       caption = "Source: Source: Stanford Open Policing Project") +
  theme_classic()

hitRatePlot
```



From this plot, we can see that Black and Hispanic drivers have a significantly less hit rate compared to the White and Asian/Pacific Islander drivers. This means that for every 100 Black drivers that were stopped and searched, less than 10 of them actually had an illegal possession while this number is over 30 for Asian Drivers. This shows that officers do not necessarily have a similar threshold (in other words, min. amount of suspicion) for drivers from any race and they employ a negative bias towards Black and Hispanic drivers in terms of searching the stopped drivers.

Binary Logistic Regression

While Xu et al. (2024) used binary logistic regression to guess if the traffic stop would be considered as legit or not, I will use binary logistic regression to see if there is a relationship between the stop resulting in an arrest and the drivers race, sex, and age. In the second part, I will examine the relationship between if the traffic stop is resulted in a search or not is related to the race, sex, and / or age of the driver.

Before starting my modeling and analysis, I created this helper function to plot the ROC and show the AUC value for our logistic regression models.

```
plotROC = function(model,
                    data,
                    responseVar,
                    figNum) {
  predicted_probs = predict(model, type = "response")
  roc_obj = roc(data[[responseVar]], predicted_probs)
  auc_value = auc(roc_obj)

  ggroc(data = roc_obj,
```

```

    size = 1.5) +
  labs(title = paste0('Fig.',
                      figNum,
                      ' ROC Curve (AUC = ',
                      round(auc_value, 3), ')'),
        subtitle = paste0('for Logistic Regression Predicting ',
                          responseVar),
        x = "Specificity",
        y = "Sensitivity") +
  theme_classic()
}

```

Following this, I can prepare my data for modeling.

```

model.data = traffic.data %>%
  filter(!is.na(subject_race), !is.na(subject_sex), !is.na(subject_age))

```

Now, I can start with the arrest model which checks if subject race, sex, and age is a significant predictor for the arrest made / not made outcome of the stop.

```

arrest_model = glm(arrest_made ~
  subject_race + subject_sex + subject_age,
  data = model.data,
  family = binomial)

summary(arrest_model)

```

```

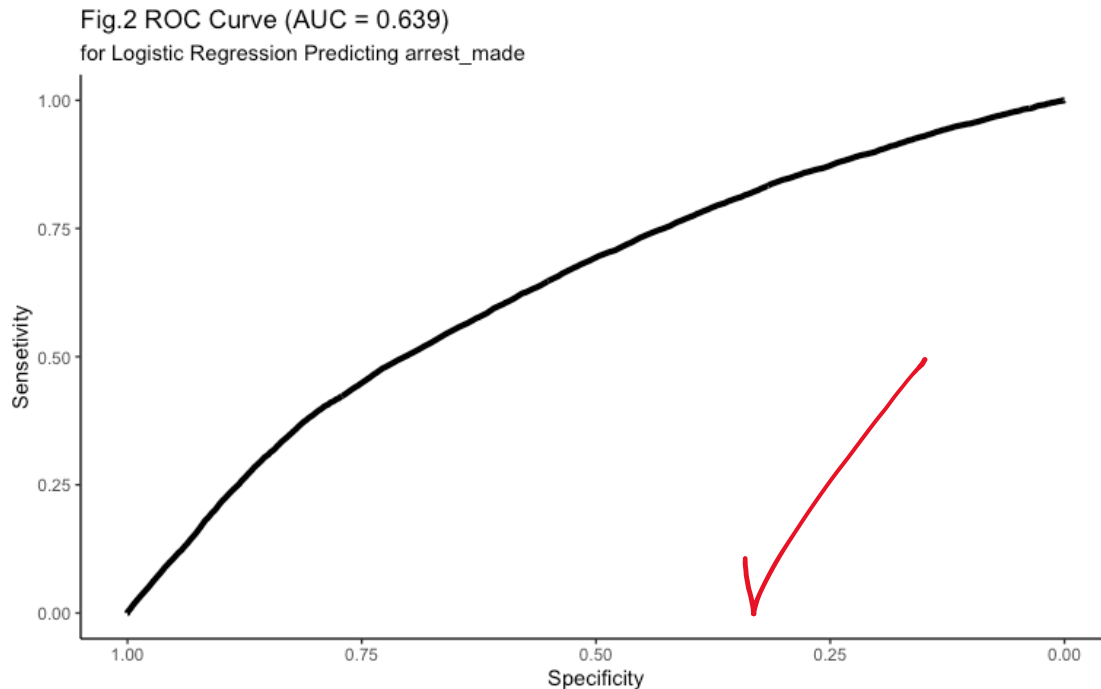
##
## Call:
## glm(formula = arrest_made ~ subject_race + subject_sex + subject_age,
##      family = binomial, data = model.data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -4.0704790   0.0413006  -98.557 <0.0000000000000002 ***
## subject_raceBlack    1.0175940   0.0348754   29.178 <0.0000000000000002 ***
## subject_raceHispanic  0.8490327   0.0371797   22.836 <0.0000000000000002 ***
## subject_raceWhite    0.3249509   0.0340696    9.538 <0.0000000000000002 ***
## subject_raceOther    0.0447622   0.0458833    0.976      0.329
## subject_sexFemale   -0.4564452   0.0237338  -19.232 <0.0000000000000002 ***
## subject_age        -0.0164734   0.0007826  -21.051 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 115529  on 805839  degrees of freedom
## Residual deviance: 112892  on 805833  degrees of freedom
## AIC: 112906
##
## Number of Fisher Scoring iterations: 7

```

I will make the interpretation of this output for Homework 8.

Following this model, I can plot the ROC curve and also see the AUC value.

```
plotROC(arrest_model, model.data, "arrest_made", 2)
```



The AUC value shows how well my model can separate these binary classes of Arrest Made and not made. An AUC value of 0.639 means that the model is able to make a correct prediction for 63.9% of the data, which is considered poor/fair prediction accuracy.

In this second part, I can check if subject race, sex, and age is a significant predictor for the search conducted / not conducted attribute of a traffic stop.

```
search_model = glm(search_conducted ~  
                    subject_race + subject_sex + subject_age,  
                    data = model.data,  
                    family = binomial)
```

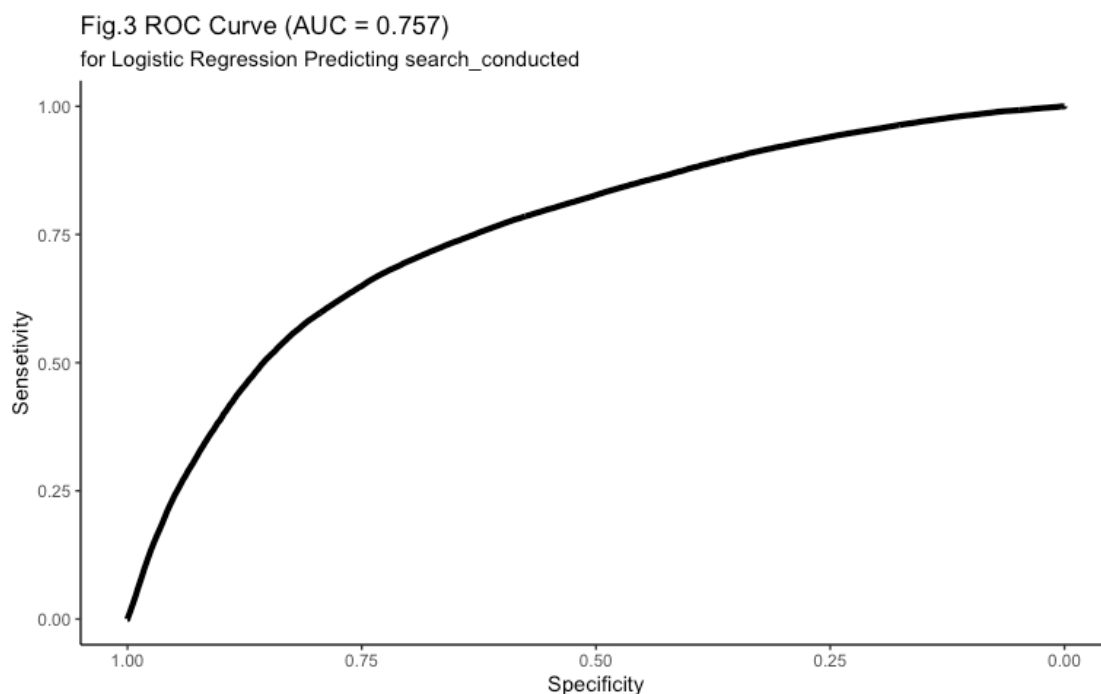
```
summary(search_model)
```

```
##  
## Call:  
## glm(formula = search_conducted ~ subject_race + subject_sex +  
##      subject_age, family = binomial, data = model.data)  
##  
## Coefficients:  
##              Estimate Std. Error z value Pr(>|z|)  
## (Intercept)   -2.5644228   0.0243909  -105.14 <0.0000000000000002 ***  
## subject_raceBlack    2.2195404   0.0211630   104.88 <0.0000000000000002 ***
```

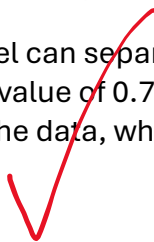
```
## subject_raceHispanic  1.6557689  0.0223232  74.17 <0.0000000000000002 ***
## subject_raceWhite     0.6043566  0.0220378  27.42 <0.0000000000000002 ***
## subject_raceOther      0.6193931  0.0263268  23.53 <0.0000000000000002 ***
## subject_sexFemale     -0.7702888  0.0126126 -61.07 <0.0000000000000002 ***
## subject_age           -0.0337612  0.0004242 -79.58 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 367945  on 805839  degrees of freedom
## Residual deviance: 326183  on 805833  degrees of freedom
## AIC: 326197
##
## Number of Fisher Scoring iterations: 6
```

I will make the interpretation of this output for Homework 8.

```
plotROC(search_model, model.data, "search_conducted", 3)
```



The AUC value shows how well my model can separate these binary classes of Search conducted and not conducted. An AUC value of 0.757 means that the model is able to make a correct prediction for 75.7% of the data, which is considered an acceptable prediction accuracy.



D.c

Explain your next step in the analysis.

As I have results for my Outcome Test, as well as the Binary Logistic regression, I will work on more of the interpretation part of the project rather than doing further EDA and modeling.

I hope that my EDA and modeling results are sufficient, so that I can write an overall report about my analysis of this dataset.

As an extra analysis step, if time permits, I am planning to try to see if there is a statistically significant relation between the location of the stop and the outcome / driver race, etc. I already have heat maps in my data analysis that can be helpful in this but I do not do any statistical analysis of this relationship. **You can also include clustering and loess**

What about Veil of Darkneess test mentioned in hw-6?

Resources

How to merge two strings in R: [StackOverflow Question](#)

General reading about the logistic regression: [Chapter 10 from R for Statistical Learning](#)

How to plot the ROC curve using ggplot: [Tutorial](#) followed

References

Pierson, E., Simoiu, C., Overgoor, J., Corbett-Davies, S., Jenson, D., Shoemaker, A., Ramachandran, V., Barghouty, P., Phillips, C., Shroff, R., & Goel, S. (2020). A large-scale analysis of racial disparities in police stops across the United States. *Nature Human Behaviour*, 4(7), 736–745. <https://doi.org/10.1038/s41562-020-0858-1>

Xu, W., Smart, M., Tilahun, N., Askari, S., Dennis, Z., Li, H., & Levinson, D. (2024). The racial composition of road users, traffic citations, and police stops. *Proceedings of the National Academy of Sciences*, 121(24). <https://doi.org/10.1073/pnas.2402547121>