Data Analysis Report

Pradyumna Das (pd10)

02/05/2021

Introduction

The history of racial bias in law enforcment in the United States is a long one. Numerous protests and riots have taken place over the years against police brutality but there hasn't been substantial change in policy. There is a significant body of literature that shows that Black and Latino populations are disproportionately impacted by violent interactions with the police. They are more likely to be over charged for small crimes, receive harsher sentences for the same crimes and have stricter conditions for bail when compared to white populations.

In the last few years there have been several high profile incidents where black people have lost their lives at the hands of the police by way of excessive use of force. This has brought the issue of racial bias in the criminal justice system to the fore again. As a result, there has been significant activity by independent researchers and journalists to gather accurate data about civilian-police interactions like stop and search etc. since the data coming out of the institutions was largely scarce.

In this study we explore such a dataset in order to analyse random searches conducted by the police to determine if there is racial bias in the selection of people that are being searched.

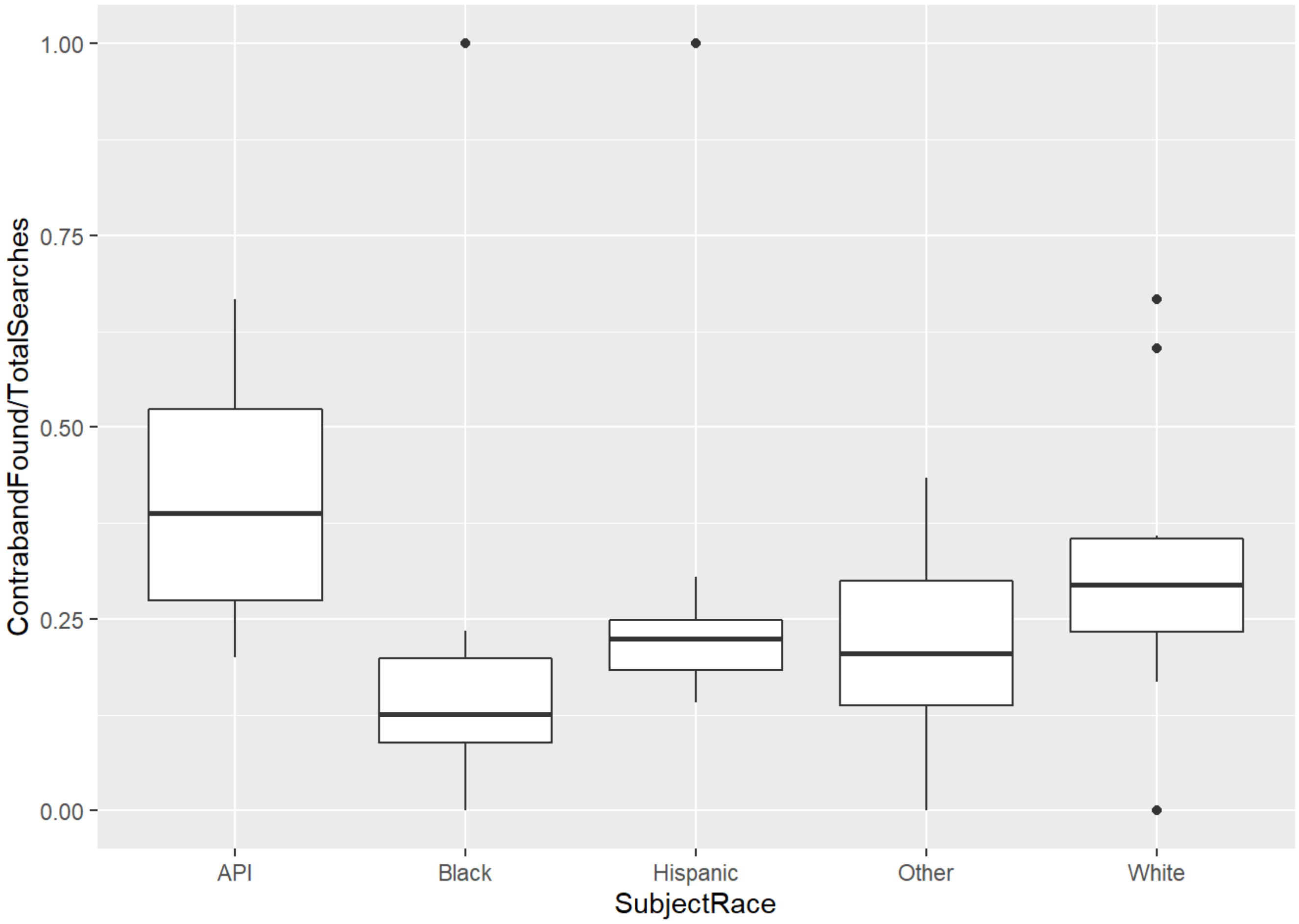
Data

We are using an aggregated version of the data used by Roberto Rivera and Janet Rosenbaum in their study titled “Racial disparities in police stops in US cities”. They in turn had taken the data from the Stanford Open Policing Project, comprised of a group of researchers studying police racial bias in most US states. The data collected by the researchers are available on their website (https://openpolicing.stanford.edu/data/).

The dataset that we will be using for this analysis contains aggregated data for searches conducted by police in San Francisco between January 1, 2015 and June 30, 2016. The dataset contains 4 columns:

1. **SubjectRace**: This column contains the race of the individuals stopped by the police. It is a discrete column with 5 values; API, Black, Hispanic, white, Other.
2. **District**: This columns contains a representation of the police district where the search occurred.
3. **ContrabandFound**: The total number of instances where the police found contraband upon the person being searched.
4. **TotalSearches**: Total number of stop and search instances that occurred

Below is a boxplot showing the spread of the raw proportions of searches, by race, that resulted in contraband being found:



- The following race/distrct combinations have no searches:

```
##   SubjectRace District
## 1         API        S
## 2         API        T
## 3    Hispanic        S
```

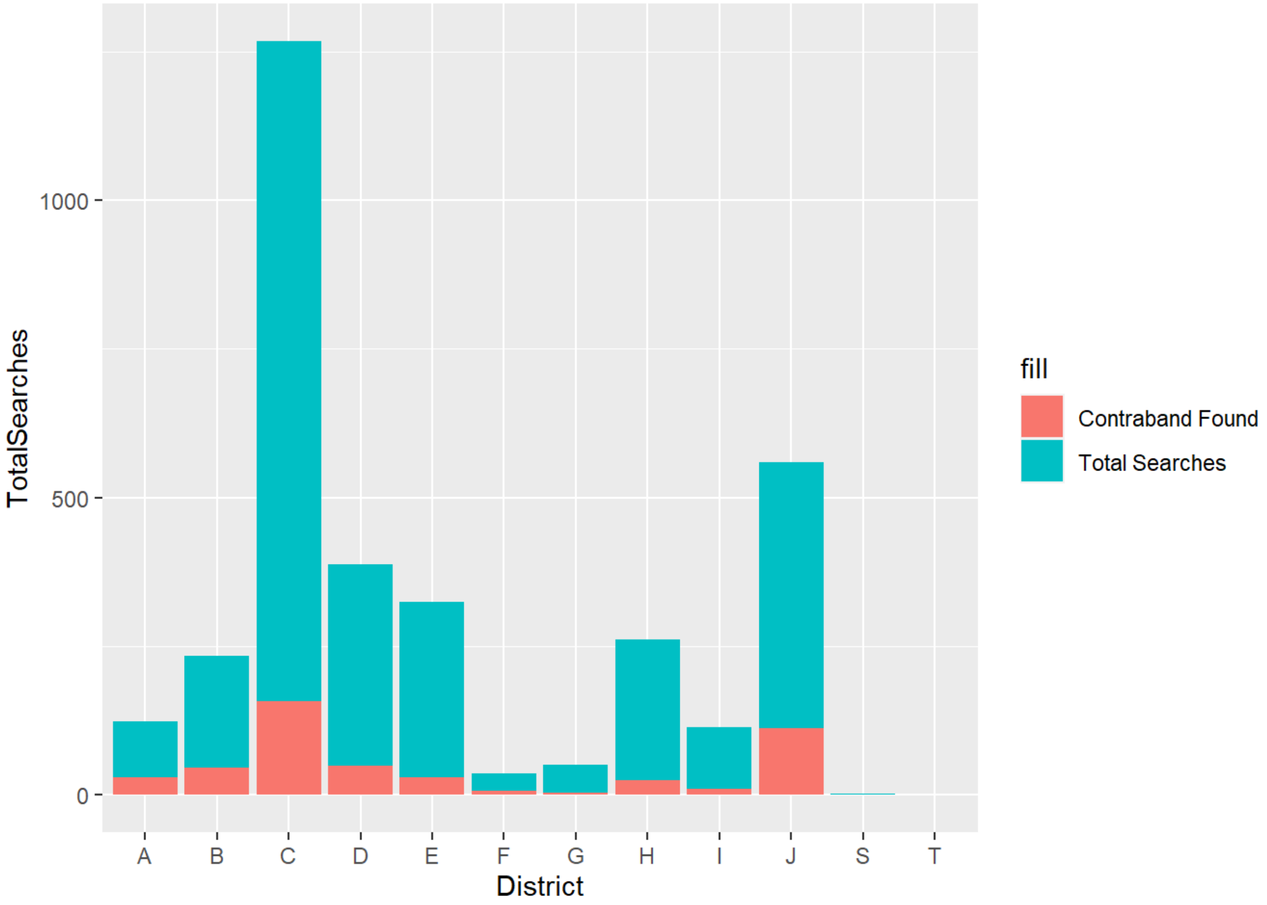
- The Asians/Pacific Islander group seems to have the highest proportion of searches that find contraband. However, we have to be wary since that group also has the highest variance stemming from the low number of total searches that happen.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 5 x 2
##   SubjectRace Searches
##   <fct>          <int>
## 1 API              258
## 2 Black           3361
## 3 Hispanic        1163
## 4 Other            349
## 5 White            967
```

Thus the result from raw proportions should not be trusted in this case.

- The black community seems to have the lowest proportion of searches that find contraband in general. The variance is also lower because there are a lot more searches (in fact its significantly more than all the other races put together). Looking at the distribution of the searches by district we can see that the searches are unevenly distributed. Nearly half the number of searches all come from district “C”.



First model

a. The first model that we use for our analysis is a logistic regression model with the following specification:

$$\begin{aligned}
&found_i \sim Bin(searches_i, p_i)\\
&logit(p_i) = \beta_i^{race} + \beta_i^{district}\\
&\beta_i^{race} \sim t(0, 10^2, 1)\\
&\beta_i^{district} \sim N(0, \sigma^2)\\
&\sigma \sim U(0, 10)
\end{aligned}$$

We model the number of instances where contraband is found on a person given the total number of searches as a binomial distribution with some probability p_i where $p_i = \frac{1}{1+e^{-(\beta_i^{race}+\beta_i^{district})}}$.

Here β^{race} and $\beta^{district}$ are the parameters of the model that control the probability. They are both random effects. β^{race} is modeled as a t-distribution and $\beta^{district}$ is modeled as a normal distribution with mean 0 and variance σ^2. σ^2 is a hyper-parameter with a uniform distribution oven the range [0, 10].

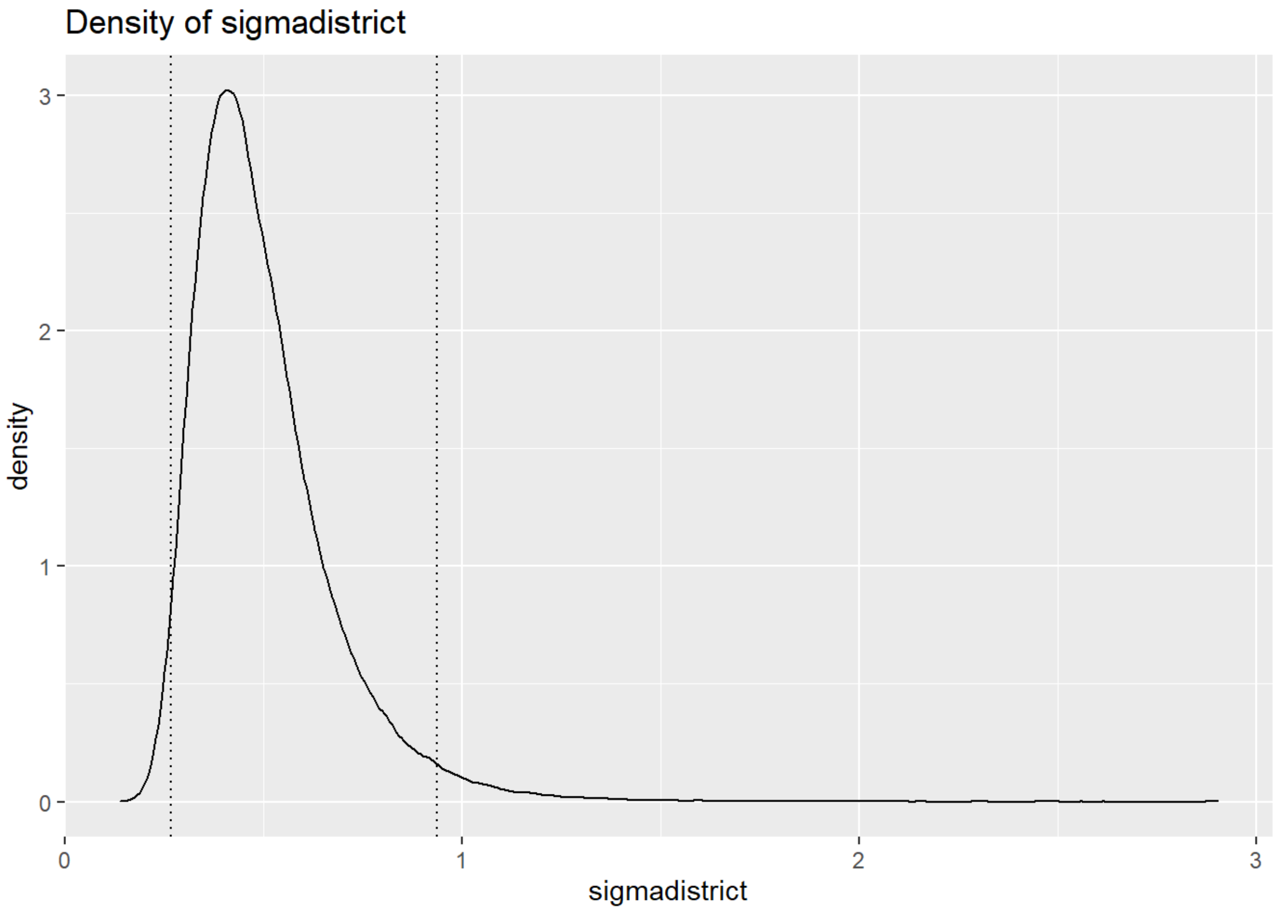
b. JAGS code for the model

```
model {

  for (i in 1:length(searches)) {
    found[i] ~ dbin(prob[i], searches[i])
    logit(prob[i]) <- betarace[race[i]] + betadistrict[district[i]]

    foundrep[i] ~ dbin(prob[i], searches[i])
  }

  for (j in 1:max(race)) {
    betarace[j] ~ dt(0, 0.01, 1)
  }

  for (k in 1:max(district)) {
    betadistrict[k] ~ dnorm(0, 1/sigmadistrict^2)
  }

  sigmadistrict ~ dunif(0,10)

}
```

c. Summary of the details of JAGS run:

- Number of chains used = 4
- Number of iterations of burn in: 15000
- Total number of samples from the posterior = 50000 * 4 = 200000
- Effective sample sizes were all above 2000
- No thinning was used

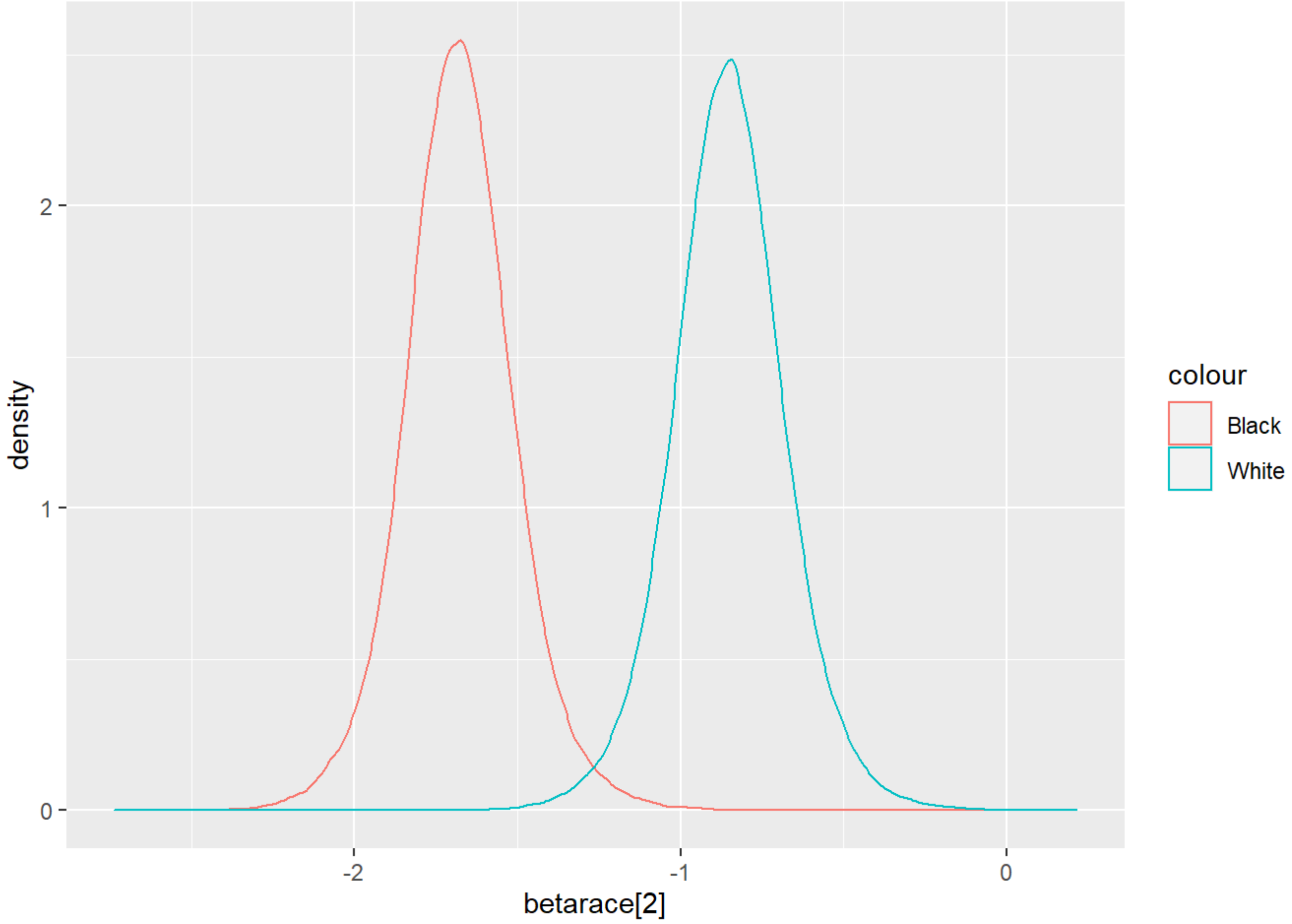
d.



As is evident from the plot, the 95% posterior confidence interval only includes positive values with a mode of around 0.4. Thus we can say with some confidence that sigmadistrict is positive which means that the different districts have different probabilities of contraband being found during a search (keeping the race constant).

e. In our model, the log odds of contraband being found on the person is modeled as the sum of β^{race} and $\beta^{district}$. Thus, keeping $\beta^{district}$ constant, if β_B is lower than β_W it means that the odds of a search unearthing contraband on a black person is lower than that of the white person.

Based on the above, we look at the posterior distributions of β_B and β_W. We find that $Pr(\beta_B < \beta_W) = 1.$ I also plotted the posterior distributions for both parameters below:



As is evident from the plot, there is almost no overlap between the parameters. We can thus conclude with confidence that the odds of a search unearthing contraband on a black person is lower than that of the white person.

f. The posterior p-value of the χ^2 statistic is 0.02452 which indicates that there is overdispersion in this model.

g. Plummer's DIC for the above model:

- Mean deviance: 298.3
- penalty 14.15
- Penalized deviance: 312.4

The effective number of parameters is around 14. This is smaller compared to the 18 parameters we have in the model (12 for the districts, 5 for race and the hyperparameter $\sigma_{district}$).

Second model

a. Code for the second model:

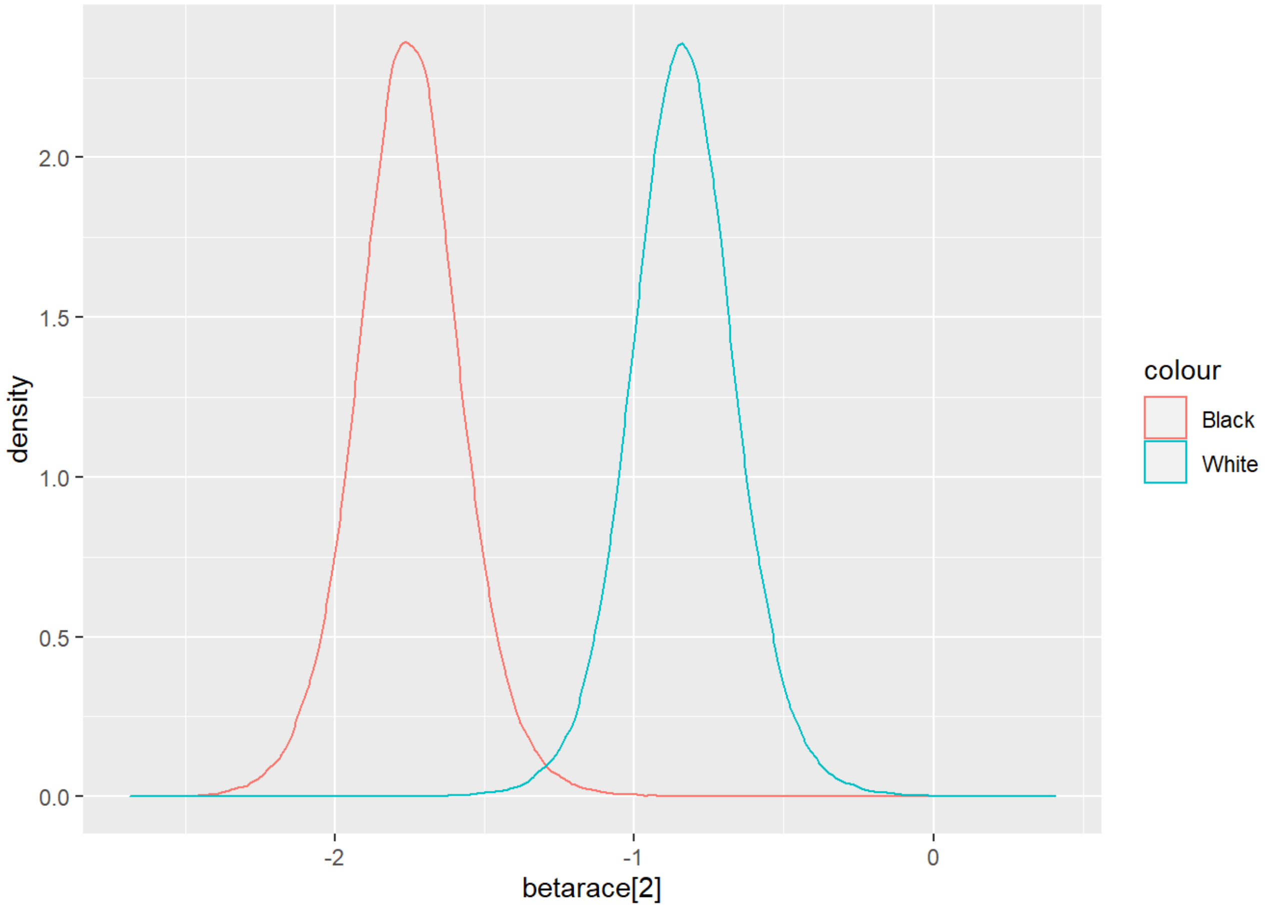
```
model {
  for (i in 1:length(searches)) {
    found[i] ~ dbin(prob[i], searches[i])
    logit(prob[i]) <- betarace[race[i]] + betadistrict[district[i]] + epsilon[i]

    epsilon[i] ~ dnorm(0, 1/sigmaepsilon^2)

    foundrep[i] ~ dbin(prob[i], searches[i])
  }

  for (j in 1:max(race)) {
    betarace[j] ~ dt(0, 0.01, 1)
  }

  for (k in 1:max(district)) {
    betadistrict[k] ~ dnorm(0, 1/sigmadistrict^2)
  }

  sigmadistrict ~ dunif(0,10)
  sigmaepsilon ~ dunif(0, 10)

}
```

b. Summary of the details of JAGS run:

- Number of chains used = 4
- Number of iterations of burn in: 25000
- Total number of samples from the posterior = 50000 * 4 = 200000
- Effective sample sizes were all well above 2000
- No thinning was used

I also calculated the posterior p-value of the χ^2 statistic as before and found that for the second model the value was 0.41783 which indicates that the problem of overdispersion was indeed fixed by addition of the normally distributed error terms.

c. The values of β_B and β_W remain largely unchanged by the addition of the random effect in the second model. The $Pr(\beta_B < \beta_W) = 1$ still. Also the posterior density plot is identical:



Thus, there is no change in our conclusion.

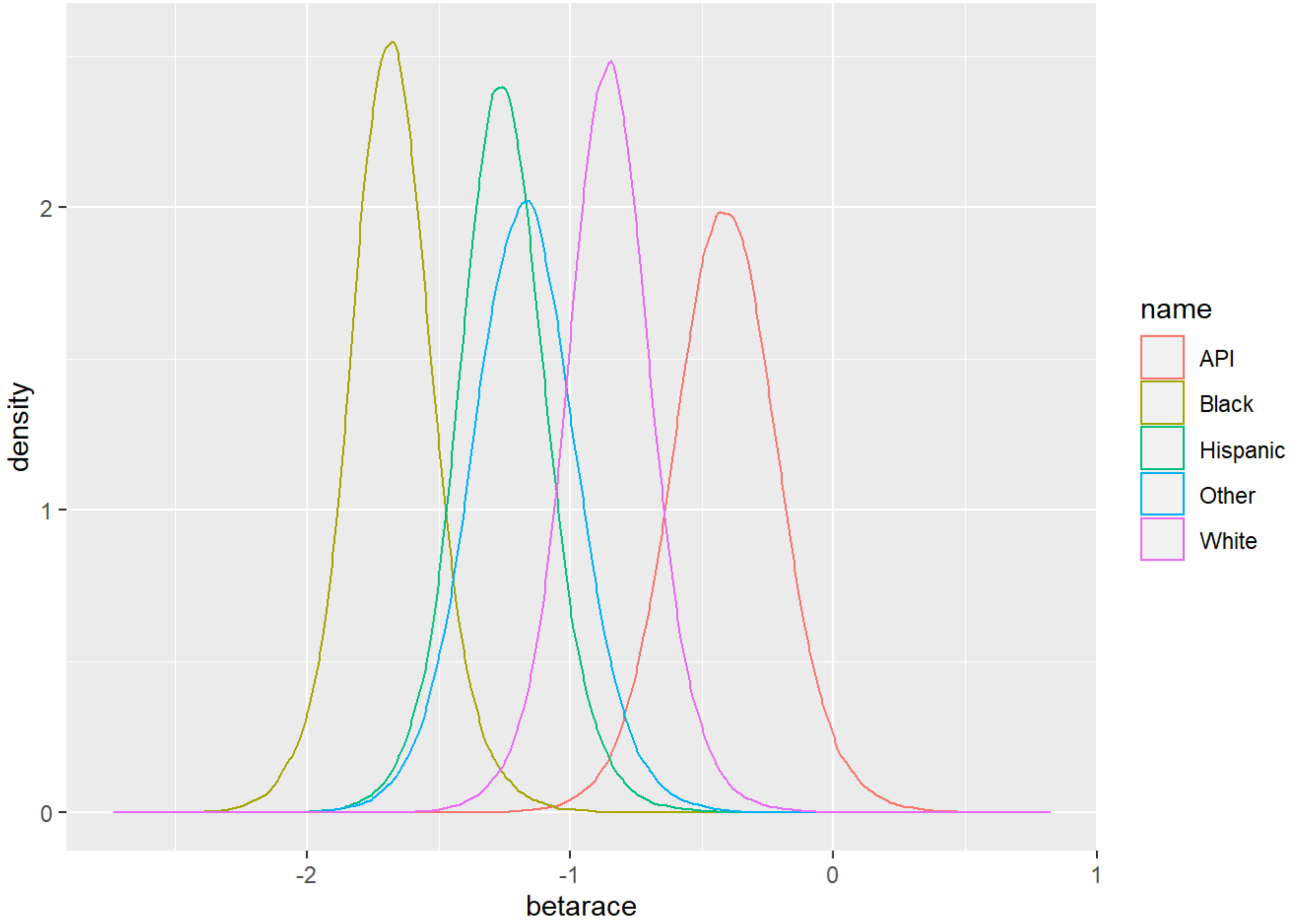
d. Plummer’s DIC:

- Mean deviance: 274.1
- penalty 28.93
- Penalized deviance: 303

The second model has a much lower deviance but twice the penalty due to the additional error terms added. Overall the model has a slightly lower DIC, making it the better than the first one.

Conclusion

In our Bayesian analysis we have tried to discern the effect of locality and race on the probability of contraband being found upon a person during a police search. We found that there indeed is an influence of both race and districts when it comes to the probability of contraband being found during a search. The posterior distributions of race parameter given a district was quite different for each of the races showing that some races are less likely to be found with contraband. Below is a plot of the β^{race} parameters:



Each of the 4 races are clearly separable in this plot, with the exception of the “Other” group which closely coincides with the hispanic population, which is interesting. Also, the probability of a black person being caught in possession of contraband is much lower than that of a white person, holding the district constant. In contrast, the raw data shows that a black person is thrice as likely as a white person to be searched (going by the raw counts of searches by race in all the districts combined). A factor of 3 seems too large to be attributed to random chance and *might* indicate a racial bias in how the searches are conducted. However, further data (e.g the demographic data of the districts) and analysis is required to reach a concrete conclusion.

Appendix

Code to setup and run the jags model:

```
# Reading in the raw data
d = read.csv("policesearchSanFrancisco.csv")

# Setting up the data to be passed to the JAGS models
jags_data = data.frame(found=d$ContrabandFound, race=unclass(d$SubjectRace), searches=d$Total
Searches, district=unclass(d$District))
jags_data = filter(jags_data, searches != 0)

# Initial values and jags setup for both the models
inits1=list(list(betarace=c(5, -5, 5, -5, 5), sigmadistrict=5),
           list(betarace=c(-5, 5, -5, 5, -5), sigmadistrict=0.01),
           list(betarace=c(5, 5, 5, -5, -5), sigmadistrict=5),
           list(betarace=c(-5, -5, 5, 5, 5), sigmadistrict=0.01))

m1 = jags.model("firstmodel.bug", jags_data, inits = inits1, n.chains = 4)

inits2=list(list(betarace=c(5, -5, 5, -5, 5), sigmadistrict=5, sigmaepsilon=5),
           list(betarace=c(-5, 5, -5, 5, -5), sigmadistrict=0.01, sigmaepsilon=5),
           list(betarace=c(5, 5, 5, -5, -5), sigmadistrict=5, sigmaepsilon=0.01),
           list(betarace=c(-5, -5, 5, 5, 5), sigmadistrict=0.01, sigmaepsilon=0.01))

m2 = jags.model("secondmodel.bug", jags_data, inits = inits2, n.chains = 4)

# Code to fetch samples from the posterior
samples1 = coda.samples(m1, c("betarace", "betadistrict", "sigmadistrict"), 50000)

# Calculating effective sample sizes
effectiveSize(samples1)

# Calculate the posterior probability of beta_B < beta_W
mean(samples1_df$`betarace[2]` < samples1_df$`betarace[5]`)
mean(samples2_df$`betarace[2]` < samples2_df$`betarace[5]`)

# Calculating Plummer's DIC
dic.samples(m1, 5000)
dic.samples(m2, 5000)

# Code for the traceplots used to check convergence
traceplot(samples1)
traceplot(samples2)

# Code for calculating the Gelman-Rubin statistic
gelman.diag(samples1)
gelman.diag(samples2)

# Code to calculate the posterior p-value for the chi-squared statistic
prob = as.matrix(rep_samples1)[, paste("prob[",1:57,"]", sep="")]
found = as.matrix(rep_samples1)[, paste("foundrep[",1:57,"]", sep="")]
chi = numeric(200000)
chirep = numeric(200000)
nonzerodata = d[d$TotalSearches > 0, ]

for(i in 1:200000){
  chi[i] = sum((nonzerodata$ContrabandFound - prob[i, ]*nonzerodata$TotalSearches)^2/(nonzero
data$TotalSearches*prob[i, ]*(1-prob[i, ])))
  chirep[i] = sum((found[i, ] - prob[i, ]*nonzerodata$TotalSearches)^2/(nonzerodata$TotalSear
ches*prob[i, ]*(1-prob[i, ]))
}
```

Code for the plots:

```
# Boxplot showing the raw proportions
ggplot(d) + geom_boxplot(mapping = aes(SubjectRace, ContrabandFound/TotalSearches)) #d is the

# Plot for the stacked bar chart showing the proportion of contraband found to searches for the black community by district
filter(d, SubjectRace=="Black") %>% ggplot() + geom_col(mapping = aes(District, TotalSearches, fill="Total Searches")) + geom_col(mapping = aes(District, ContrabandFound, fill="Contraband Found"))

# Plot for the density of sigmadistrict. samples1_df is just the posterior samples from coda.samples in a DF
ggplot(samples1_df) + geom_density(aes(sigmadistrict)) + geom_vline(xintercept = quantile(samples1_df$sigmadistrict, c(0.025, 0.975)), linetype=3) + ggtitle("Density of sigmadistrict")

# Plot for comparing the posterior densities of beta_W and beta_B
ggplot(samples1_df) + geom_density(aes(`betarace[2]`, colour="Black")) + geom_density(aes(`betarace[5]`, colour="White"))

#Code to generate the race-wise density plots in the conclusion
racedata = samples1_df[, paste("betarace[",1:5, "]", sep="")]
names(racedata) = c("API", "Black", "Hispanic", "Other", "White")
lrd = pivot_longer(racedata, everything())
ggplot(lrd) + geom_density(aes(value, colour=name))
```

Citations

1. E. Pierson, C. Simoiu, J. Overgoor, S. Corbett-Davies, D. Jenson, A. Shoemaker, V. Ramachandran, P. Barghouty, C. Phillips, R. Shroff, and S. Goel. “A large-scale analysis of racial disparities in police stops across the United States”. Nature Human Behaviour, Vol. 4, 2020.

2. Racial bias in criminal news in the United States (https://en.wikipedia.org/wiki/Racial_bias_in_criminal_news_in_the_United_States)

3. 100 years of racism in policing (https://www.aclu.org/news/criminal-law-reform/what-100-years-of-history-tells-us-about-racism-in-policing/)

4. Stop and frisk policy in New York (https://en.wikipedia.org/wiki/Stop-and-frisk_in_New_York_City)