Final Project: New York City Stop-Question-Frisk Group Write Up

Xingyao Chen, Sonia Sehra, Dave Makhervaks, Jenna Kahn Section 5

Due December 13, 2016

Background provided by instructors:

In stop-question-frisk, a police officer is authorized to stop a pedestrian, question them, and then frisk their body searching for contraband items such as weapons or drugs. The motivation for the policy was to prevent crimes from happening in the first place, though recent studies by the New York Civil Liberties Union suggest the policy did not achieve noticeable reductions in crime. For example, NYCLU estimates that guns were found in fewer than 0.2% of stops. Moreover, the policy was found to be discriminatory because Blacks and Latinos were stopped disproportionately more than their participation in crime would suggest. (Since 2014, new restrictions have limited the use of SQF, and the numbers of SQF incidents have decreased precipitously.)

Data from New York City's stop-question-frisk (SQF) program is publicly available online. Whenever a person is stopped under SQF, the officer is required to fill out a form with information about the stop. Each year, the data is compiled and released by the city.

Required Questions:

i. Because all of the data provided are reported by the police officers who stopped the suspects, it is possible that there is misreported data. For example, an officer may misidentify an Asian/Pacific Islander as Hispanic, or vice-versa. We would like to explore what effect such misidentifications can have on the collected data. To simplify our analysis, we will only consider the subset of the data for which suspects were identified as Asian/Pacific Island or Hispanic. Suppose that a Hispanic person has a 95% chance of correctly being identified as Hispanic, and otherwise is misidentified as Asian/Pacific Islander, and an Asian/Pacific Islander has a 95% chance of correctly being identified as Asian/Pacific Islander, and otherwise is misidentified as Hispanic. Using this subset of the 2010 data, determine the probability that a stopped suspect is Hispanic and the probability that a stopped suspect is Asian/Pacific Islander. Use this to determine the probability that someone who is identified by the officer as Hispanic is actually Hispanic and the probability that someone who is identified as Asian/Pacific Islander is actually Asian/Pacific Islander. What does this say about how we utilize the `race` column of the data?

Refer to R code at *Question 1 Code* in "Group Code and Graphs.rmd"

LH=suspect is labeled as Hispanic

LA=suspect is labeled as Asian

A=suspect is Asian

H=suspect is Hispanic

$$\begin{aligned} P(LH) &= P(LH|H)P(H) + P(LH|A)P(A) \\ P(LA) &= P(LA|H)P(H) + P(LA|A)P(A) \\ 0.9 &= 0.95P(H) + 0.05P(A) \\ 0.09 &= 0.05P(H) + 0.95 * P(A) \end{aligned}$$

$$\begin{aligned} P(H) &= 0.95067 \\ P(A) &= 0.04933 \end{aligned}$$

The conditional probabilities

$$
\begin{aligned}
P(H|LH) &= \frac{P(LH|H)P(H)}{P(LH)} \\
P(A|LA) &= \frac{P(LA|A)P(A)}{P(LA)} \\
P(H|LH) &= 0.997 \\
P(A|LA) &= 0.4964
\end{aligned}
$$

ii. Compare the rates at which suspects stopped in 2010 were frisked, broken down by race. Are the differences in rates between the various groups statistically significant? Which borough(s) had the largest differences? The smallest?

Refer to R code and table at *Question 2 Chunk 1 Code* in "Group Code and Graphs.rmd"

The largest difference is between BRONX and BROOKLYN, BRONX and BROOKLYN, BRONX and MANHATTAN, BRONX and STATEN IS, BROOKLYN and QUEENS, MANHATTAN and QUEENS, and QUEENS and STATEN IS.

Refer to R code and table at *Question 2 Chunk 2 Code* in "Group Code and Graphs.rmd"

The smallest difference is between BROOKLYN and MANHATTAN, UNKNOW and BROOKLYN, and UNKNOW and STATEN IS, UNKNOW and MANHATTAN.

iii. Compare the distribution of ages for male suspects in 2010 with the distribution of ages for female suspects in 2010. Use `qqnorm` to determine if they are normally distributed, and compare them with each other by using `qqplot`. Are the age distributions the same? Compare the age distribution for male suspects to that of the entire population of suspects that were stopped in 2010. Type `?qqplot` for help on how to use the command. Note that some ages are reported as 0 or 999 if the officer did not know the age, so you may want to throw out the extraneous data first.

Refer to R code and figures at *Question 3 Code* in "Group Code and Graphs.rmd"

The distribution of ages for male suspects with the outliers included is somewhat normal in the middle, but it is pretty messy and is not a very good QQ plot (not a straight line). The same follows for the distribution of ages for female suspects with the outliers included in the data. However, when the outliers are removed, then the QQ-Plot for both the distribution of both male and female ages are reasonably normal. We know this, because the QQ plots without the outliers are reasonably straight linear excluding the sides of the graph. Creating a QQ Plot of Male ages vs Female ages, we get a very linear line, showing that the male and female distributions are in fact, very similar! Comparing the male distribution of ages against the age distribution of the entire population also yields a linear plot, showing that the distributions are very similar! The distributions between Females and Males are approximately the same, because the `qqplot` results yielded points that fall pretty closely on the $y = x$ line.

iv. Find the probability, with a 95% confidence interval, that a suspect was frisked (a) for the entire population in 2015, and (b) for suspects in 2015 who refused to provide identification, and determine whether suspects who refused to provide identification had a different probability of being frisked than the population at large.

(a) Refer to R code at *Question 4 Chunk 1 Code* in "Group Code and Graphs.rmd"

The probability, with a 95% confidence interval, that a suspect was frisked for the entire population in 2015 is 0.6761955 ± 0.006105832.

(b) Refer to R code at *Question 4 Chunk 2 Code* in "Group Code and Graphs.rmd"

The probability, with a 95% confidence interval, that a suspect who refused to provide ID was frisked is 0.6103286 ± 0.03784224.

Refer to R code at *Question 4 Chunk 3 Code* in “Group Code and Graphs.rmd”

A two sample t-test can be used to determine whether these rates significantly differ. The t-test suggests there is a significant difference in the probability a suspect in general will be frisked vs. the probability a suspect who refuses to provide ID will be frisked ($p < .05$).

v. For the 2010 data, decide which of the following binary factors: `arstmade`, `searched`, `inside`, `sumissue`, `frisked`, `weap`, `contrabn radio`, `pf` had a significant effect on the length of the stop (`perstop`) by using linear regression. Make sure to check your residuals for normality, and apply an appropriate transformation to `perstop` or remove outlier points if it does not look normal (see your notes from Lecture 11 to review how to do this). Note that `perstop` is a discrete variable, so you are looking for an approximately normal distribution for the residuals. Consider the p-values for the coefficients and the R^2 value for your regression model. What do they indicate about how the factors affect the length of the stop? Recall that $R^2 = \frac{SSR}{SST}$. How much of the variability in `perstop` is due to the explanatory variables you have selected? Why does this make sense?

Refer to R code and tables at *Question 5 Code* in “Group Code and Graphs.rmd”

The R^2 values indicate that the linear models are a poor fit for the data. Because of the low R^2 values, not much variability is explained by the selected variables.

To answer this question, it is useful to understand how to interpret a regression model involving indicator variables. Suppose you have an indicator variable X that equals 1 when a condition is true and 0 otherwise, and you fit the following regression model:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Note that when $X = 0$, meaning that the condition is false, $Y = \beta_0 + \epsilon$, and when $X = 1$, meaning that the condition is true, $Y = \beta_0 + \beta_1 + \epsilon$. Therefore, β_1, the coefficient on X in the regression model, gives the *average effect* that the condition has on the outcome. That is β_1 equals the mean difference in the response when the condition is true compared to when it is false. Our hypothesis is constructed in the same way as always: $H_0 : \beta_1 = 0$ is the null hypothesis asserting that the condition has no effect on Y. $H_a : \beta_1 \neq 0$ is the alternative hypothesis asserting that the condition has some effect on Y. If the p-value on β_1 after we fit the regression model is smaller than our significance level α, then we reject H_0 and conclude that the condition being true has a statistically significant effect on the value of Y.

Team-Chosen Questions:

Introduction:

We examined whether people of certain races are more likely to be stopped and/or frisked than would be expected based on the demographic makeup of each borough. This question is important because stop and frisk has faced numerous accusations of racial profiling. If we find that certain racial groups are disproportionately likely to be stopped that, then our findings give support to the claim that stop and frisk unfairly targets those groups. To answer our question, we examined two sets of data. One is the 2010 US government census data for each of the five New York boroughs. We used this find the racial breakdown of population of each borough. The other data set is a collection of police filings from all of the stop-and-frisk incidents in New York in 2010. Each filing contains information on the race of the stopped individual, where the stop occurred, and whether the individual was frisked during the encounter. We used this data set to the relative frequency with which people of each race were likely to be stopped and frisked. We conducted graphical and statistical analyses to determine whether black, Hispanic, and white people were disproportionately likely to be stopped and frisked

Question 1: Are people stopped in proportion to the demographic characteristics of the area?

Refer to R code, figures and tables at *Team-Chosen Question 1 Chunks 1 and 2 Code* in “Group Code and Graphs.rmd”

Question 2: Are they frisked in proportion to the demographic characteristics of the area?

Refer to R code, figures and tables at *Team-Chosen Question 2 Chunks 1 and 2 Code* in “Group Code and Graphs.rmd”

Analysis

In order to test if the difference between the percentage of people of a certain race stopped and the percentage of people of a certain race who live in that borough are drastically different, we used the `prop.test` function in R. For both the black and Hispanic populations we said that our alternative hypothesis was that the percentage of stopped people is greater than the percentage of people living in that borough. We found that the p-value calculated using this function gave us roughly zero for the percentage of stopped people who are black versus the percentage of black people in that borough. Since the p-value is less than 0.05 than we can confidently say that these two percentages are significantly different (which is a problem). The H_0 is rejected. Performing the same prop.test for the Hispanic population of each other five boroughs, we get practically identical results, all the p-values are practically 0 (which is less than 0.05) so we can confidently say that these two percentages are significantly different for Hispanics as well. As for the white population, we said the alternative hypothesis was that the percentage of stopped people who are white is less than the percentage of white people in that borough. The p-values we got for each of the five boroughs was approximately 0, rejecting our H_0. This means that we can say that the percentage of white people in the borough is significantly greater than the percentage of stopped people that are white.

We also did the same prop.test for the percentage of people frisked in the borough versus percentage of people living in the borough. For the African American population we found that the p-value also approximated to 0 for each of the black boroughs for the some hypothesis we had above. Our results were also identical to the ones above for both Hispanics and Whites.

Concluding Remarks

After looking at all this data, we can make a few important conclusions. First, in New York City and most likely the rest of the United States, our police departments do racial profiling and stop and frisk an disproportionate amount of African American and Hispanic people. Second, there must be changes made to this problem. Some proposed changes could be a decrease in the amount of stop and frisk that police departments around the nation and more particularly, in NYC, do. In addition, racial profiling must be addressed in the education and training of police officers around the United States in order to ensure equal treatment of all citizens and residents of this country.

Bibliography

Barone, M., “Stop-and-Frisk Protects Minorities”, http://www.nationalreview.com/article/356481/stop-and-frisk-protects-minorities-michael-barone, 2013.

New York Civil Liberties Union website: http://www.nyclu.org/node/1598

New York Civil Liberties Union, “Stop and Frisk During the Bloomberg Administration”, http://www.nyclu.org/files/publications/stopandfrisk_briefer_2002-2013_final.pdf

Geller, A., Fagan, J., Tyler, T., Link, B., “Aggressive Policing and the Mental Health of Young Urban Men”, *Am J Public Health.* 2014 December; 104(12): 2321-2327. Published online 2014 December. doi: 10.2105/AJPH.2014.302046