U6614: Subway Fare Evasion Arrests and Racial Bias

Your Name (your-uni)

2025-02-08

Please submit your knitted .pdf file along with the corresponding R markdown (.rmd) via Courseworks by 11:59pm on the due date.

1 Load libraries

2 Aggregating to subway station-level arrest totals

2a) Load full set of cleaned arrest microdata (arrests.clean.rdata).

2b) Create new data frame (`st_arrests`) that aggregates the microdata to station-level observations including columns for `st_id`, `loc2` and `total arrests`:

2c) Plot histogram of arrests and briefly describe the distribution of arrests across stations.

3 Joining subway ridership and neighborhood demographic data

3a) Read in poverty and ridership csv files with strings as factors (`station_povdataclean_2016.csv` and `Subway Ridership by Station - BK.csv`).

3b) Join both data frames from 3a to `st_arrests` and inspect results (store new data frame as `st_joined`).

- Inspect results from joins, drop unnecessary columns from the ridership data, and group `st_joined` by `st_id` and `mta_name`.
- Only display ungrouped version of `st_joined` for compactness.

3c) Print the top 10 stations by total arrest counts

- Only display `st_id`, `mta_name`, `arrests_all`, `shareblack`, `povrt_all_2016` (no other columns)

4 Explore relationship between arrest intensity and poverty rates across subway station (areas)

4a) Compute arrest intensity and other explanatory variables for analysis.

- Drop the observation for the Coney Island station and very briefly explain your logic
- Create new column of data for the following:
  - fare evasion arrest intensity: `arrperswipe_2016` = arrests per 100,000 ridership ('swipes')
  - a dummy indicating if a station is high poverty: `highpov` = 1 if pov rate is > median pov rate across all Brooklyn station areas
  - a dummy for majority Black station areas: `nblack` = 1 if `shareblack` > 0.5
- Coerce new dummy variables into factors with category labels
- Assign results to new data frame called `stations`
- Display top 10 station areas by arrest intensity using `kable()` in the `knitr` package

4b) Examine the relationship between arrest intensity and poverty rates

- Show a scatterplot of arrest intensity vs. poverty rates along with the regression line you think best fits this relationship.
- Which regression specification do you prefer: linear or quadratic? Be clear about your logic and if applicable cite statistical evidence to support your decision.
- Explain your logic about whether to weight observations or not.
- Interpret your preferred regression specification (carefully!).

4c) Estimate and test the difference in mean arrest intensity between high/low poverty areas

- Report difference and assess statistical significance
- Weight observations by ridership

5 How does neighborhood racial composition mediate the relationship between poverty and arrest intensity?

- In this section, you will examine the relationship between arrest intensity & poverty by Black vs. non-Black station area (`nblack`).

5a) Present a table showing the difference in mean arrests intensity for each group in a 2x2 table of `highpov` vs `nblack`.

- Remember to weight by ridership at each station
- Could the difference in arrest intensity be explained by differences in poverty rate?

5b) Show a scatterplot of arrest intensity vs. poverty rates (with separate aesthetics for Black and non-Black station areas) along with the regression line you think best fits this relationship.

- Which regression specification do you prefer: linear or quadratic? Be clear about your logic and if applicable cite statistical evidence to support your decision.
- Interpret your preferred regression specification (carefully!).

5c) Next let's let's think about how measurement error might impact results from 5b. Do you think measurement error could bias your estimates of neighborhood racial gaps in the effect of poverty on enforcement intensity from 5b? Explain, carefully. Do you have any creative ideas to address any concerns you have about potential bias due to measurement error?

- One source of measurement error owes to the fact that we're using racial-ethnic composition and poverty rates for the neighborhood surrounding each station to proxy for characteristics of riders at each station. These variables are measured with *non-random* error; demographic measures for the surrounding neighborhood will tend to be a less accurate proxy for the demographics of riders at that station for busier stations that are destinations for commuters, tourists and others who may not live in very vicinity close to the station.
- Tip: this is a very tricky issue! In order to think through the measurement error problem and it's consequences you will probably want to consult your Quant II notes and/or my Quant II video lecture 4 on the course website.
- Can you think of any other measurement error problems that might affect your results from 5b?
- Do you have any creative ideas for addressing any concerns you have about potential bias due to this source of measurement error, using this data or other data you think might exist?

6 Examine the relationship between arrest intensity and crime

6a) Load the crime data (`nypd_criminalcomplaints_2016.csv`) and join to the existing `stations` data frame.

NOTE: For the next two subsections, present your preferred plots to inform the relationships in question, along with any additional data manipulation and evidence to support your decisions/interpretation/conclusions. You'll want to explore the data before arriving at your preferred plots, but don't present everything you tried along the way such as intermediate versions of your preferred plot. Focus on the analysis you eventually settled on to best inform the question at hand, and any critical observations that led you down this path.

6b) Examine the overall relationship between arrest intensity and crime (without taking neighborhood racial composition or poverty into account) (comparable to Section 4b). Carefully interpret the results you choose to present.

6c) Examine how neighborhood racial composition mediates the relationship between arrest intensity and crime (comparable to Section 5b). Carefully interpret the results you choose to present.

7 Summarize and interpret your findings with respect to subway fare evasion enforcement bias based on race

- Is there any additional analysis you'd like to explore with the data at hand?
- Are there any key limitations to the data and/or analysis affecting your ability to assess enforcement bias based on race?
- Is there any additional data you'd like to see that would help strengthen your analysis and interpretation?
- For this question, try to be specific and avoid vaguely worded concerns.