U6614: Subway Fare Evasion Arrests and Racial Bias

Sample Solution

2025-02-12

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

```
library(tidyverse)
library(weights)
library(lmtest)
library(sandwich)
library(knitr)
library(estimatr)
library(ggpmisc)
```

2 Aggregating to subway station-level arrest totals

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

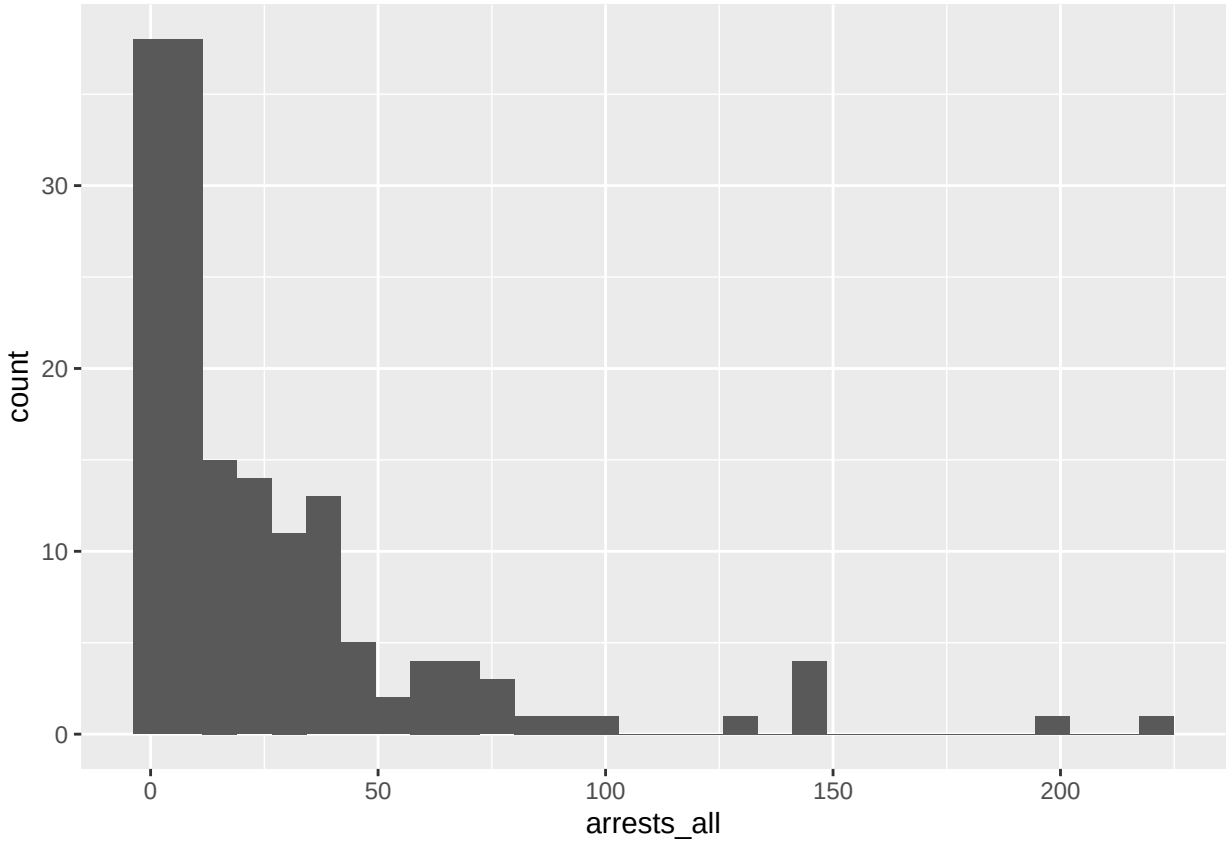
```
load("arrests.clean.RData")
```

2b) Using tidyverse functions, create a new data frame (`st_arrests`) that aggregates the microdata to station-level observations. For `st_arrests`, the unit of analysis should be the station, with columns for `st_id`, `loc2` and `total arrests`.

```
st_arrests <- arrests.clean %>%
  group_by(st_id, loc2) %>%
  summarise(arrests_all = n() ) %>%
  arrange(desc(arrests_all))
```

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

```
ggplot(data = st_arrests, aes(x = arrests_all)) +
  geom_histogram()
```



This histogram shows that the majority of subway stations had a relatively small number of fare evasion arrests. The median station arrest total is 13 compared to a mean of 26.82, with 8 stations home to more than 100 arrests.

3 Joining subway ridership and neighborhood demographic data and prepping data for analysis.

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

```
st_poverty <- read.csv("station_povdataclean_2016.csv",
                       stringsAsFactors = TRUE)

st_ridership <- read.csv("Subway Ridership by Station - BK.csv",
                         stringsAsFactors = TRUE)
```

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 ridership columns ("swipes") from the ridership data, and group `st_joined` by `st_id` and `mta_name`.
- Only display ungrouped version of `st_joined` for compactness.

```
drop_vars <- c("swipes2011", "swipes2012", "swipes2013", "swipes2014", "swipes2015")

st_arrests <- st_arrests %>%
  mutate(st_id = as.integer(st_id))

st_joined <- st_arrests %>%
  inner_join(st_poverty, by = c("st_id")) %>%
  inner_join(st_ridership, by = c("st_id" = "st_id",
                                  "mta_name" = "mta_name")) %>%
  select(-all_of(drop_vars)) %>%
  group_by(st_id, mta_name)

# display structure of ungrouped data frame to avoid lengthy output listing every group
  st_joined %>% ungroup() %>% str()
```

```
## tibble [157 x 14] (S3: tbl_df/tbl/data.frame)
##  $ st_id         : int [1:157] 66 99 150 70 114 131 54 147 106 123 ...
##  $ loc2          : Factor w/ 157 levels "15 st prospect park f g line",..: 66 100 149 148 110 129 54
##  $ arrests_all   : int [1:157] 223 198 143 142 141 141 133 102 90 86 ...
##  $ x             : num [1:157] -74 -74 -73.9 -73.9 -74 ...
##  $ y             : num [1:157] 40.6 40.7 40.7 40.7 40.7 ...
##  $ mta_name      : Factor w/ 157 levels "15 St-Prospect Park F subway G subway",..: 66 99 150 70 114
##  $ pop_black_2016: int [1:157] 36 1939 14825 13135 1542 10311 5624 11804 16176 2698 ...
##  $ pov_black_2016: int [1:157] 2 677 4592 3796 483 2437 900 6706 3832 306 ...
##  $ pop_all_2016  : int [1:157] 5186 12437 18556 17561 23711 15934 6753 15751 20610 13654 ...
##  $ pov_all_2016  : int [1:157] 1329 1939 6149 5565 9182 3511 1156 9104 4809 1221 ...
##  $ povrt_all_2016: num [1:157] 0.256 0.156 0.331 0.317 0.387 ...
##  $ shareblack    : num [1:157] 0.00694 0.15591 0.79893 0.74796 0.06503 ...
##  $ nblack        : int [1:157] 0 0 1 1 0 1 1 1 1 0 ...
##  $ swipes2016    : int [1:157] 5025598 13091255 5152649 9051970 4272443 5861658 3897784 1435112 2031
```

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). Round percentages to 2 decimal points for this question and all subsequent questions.

- For this and subsequent tables, we recommend passing your table into the `kable()` function to improve the appearance. Just add `%>% kable()` at the end of your pipe.

```
st_joined %>%
  arrange(desc(arrests_all)) %>%
  select(st_id, mta_name, arrests_all, shareblack, povrt_all_2016) %>%
  mutate(shareblack = round(shareblack, 2),
         povrt_all_2016 = round(povrt_all_2016, 2)) %>%
  head(n = 10) %>%
  kable()
```

st_id	mta_name	arrests_all	shareblack	povrt_all_2016
66	Coney Island-Stillwell Av D subway F subway N subway Q subway	223	0.01	0.26
99	Jay St-MetroTech A subway C subway F subway R subway	198	0.16	0.16
150	Utica Av A subway C subway	143	0.80	0.33
70	Crown Heights-Utica Av 3 subway 4 subway	142	0.75	0.32
114	Marcy Av J subway M subway Z subway	141	0.07	0.39
131	Nostrand Av A subway C subway	141	0.65	0.22
54	Canarsie-Rockaway Pkwy L subway	133	0.83	0.17
147	Sutter Av L subway	102	0.75	0.58
106	Kingston-Throop Avs C subway	90	0.78	0.23
123	Nevins St 2 subway 3 subway 4 subway 5 subway	86	0.20	0.09

3d) 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 stations by arrest intensity using `kable()` in the `knitr` package

```
stations <- st_joined %>%
  filter(st_id != 66) %>%
  mutate(arrperswipe = round(arrests_all / (swipes2016 / 100000), 2),
         highpov = as.numeric(povrt_all_2016 > median(st_joined$povrt_all_2016)),
         nblack = as.numeric(shareblack > .5),
         highpov = factor(highpov, levels = c(0,1),
                          labels = c("Not high poverty", "High poverty")),
         nblack = factor(nblack, levels = c(0,1),
                         labels = c("Majority non-Black", "Majority Black")),
         shareblack = round(shareblack, 2),
         povrt_all_2016 = round(povrt_all_2016, 2))
```

```r
#display top 10 stations by arrest intensity (show st_id, mta_name, arrests_all and new variables)
  stations_top10 <- stations %>%
    arrange(desc(arrperswipe)) %>%
    select(st_id, mta_name, arrperswipe, arrests_all, shareblack,
           povrt_all_2016, highpov, nblack) %>%
    head(n = 10)
  kable(stations_top10)
```

st_id	mta_name	arrperswipe	arrests_all	shareblack	povrt_all_2016	highpov	nblack
101	Junius St 3 subway	11.00	75	0.78	0.48	High poverty	Majority Black
26	Atlantic Av L subway	8.48	37	0.66	0.51	High poverty	Majority Black
111	Livonia Av L subway	7.17	75	0.83	0.45	High poverty	Majority Black
147	Sutter Av L subway	7.11	102	0.75	0.58	High poverty	Majority Black
106	Kingston-Throop Avs C subway	4.43	90	0.78	0.23	High poverty	Majority Black
112	Lorimer St J subway M subway	4.39	70	0.15	0.34	High poverty	Majority non-Black
140	Rockaway Av 3 subway	3.97	61	0.78	0.40	High poverty	Majority Black
54	Canarsie-Rockaway Pkwy L subway	3.41	133	0.83	0.17	Not high poverty	Majority Black
141	Rockaway Av C subway	3.41	61	0.80	0.22	Not high poverty	Majority Black
144	Shepherd Av C subway	3.40	36	0.61	0.30	High poverty	Majority Black

3e) How do the top 10 stations by arrest intensity compare to the top 10 stations by arrest count?

Only 3 of the top 10 stations by arrest count are also in the top 10 according to arrest intensity. This demonstrates the importance of measuring arrests relative to ridership.

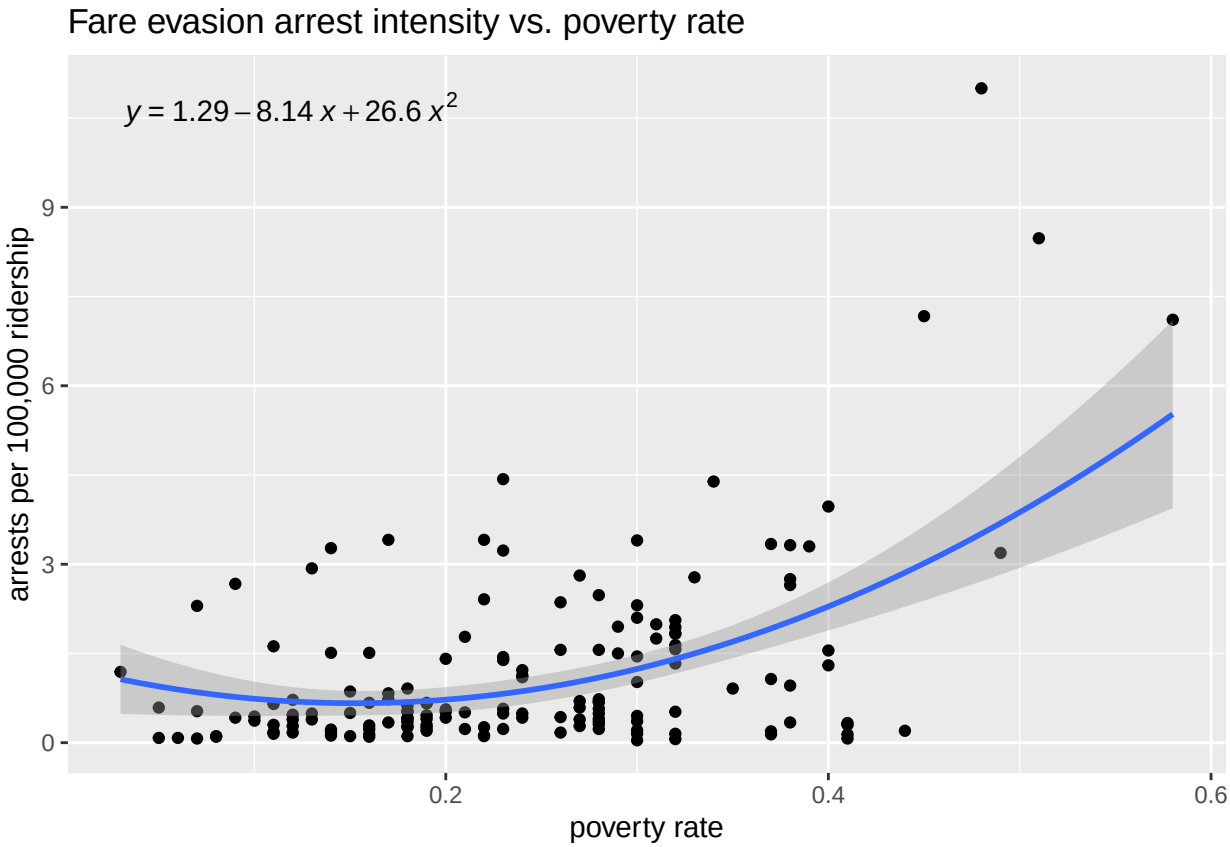
4 Explore relationship between arrest intensity and poverty rates across subway station areas.e

4a) 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. Weight observations by ridership, and label your axes appropriately. **Only show one plot with your preferred specification!**
- Which regression specification do you prefer: linear or quadratic? Be clear about your logic and cite statistical evidence to support your decision.
- Interpret your preferred regression specification (carefully!). Remember to test for statistical significance for any estimates you choose to emphasize.

```
# specify quadratic formula to refer back to
  formula <-  y ~ poly(x, 2, raw = TRUE)

  ggplot(stations, #specify data frame to use
         aes(x = povrt_all_2016,
             y = arrperswipe,
             weight = swipes2016)) + #specify columns to use
    geom_point() + #specify plot geometry
    ggtitle('Fare evasion arrest intensity vs. poverty rate') + #add title
    labs(x = 'poverty rate',
         y = 'arrests per 100,000 ridership') + #change axis labels
    geom_smooth(method = 'lm',
                formula = formula) +  #add regression line
    stat_poly_eq(mapping = use_label("eq"),
                 formula = formula)
```




```
#linear model (all stations)
ols1l <- lm_robust(arrperswipe ~ povrt_all_2016,
                   data = stations,
                   weights = swipes2016)
#summary(ols1l)

#quadratic model(all stations)
ols1q <- lm_robust(arrperswipe ~ povrt_all_2016 + I(povrt_all_2016^2),
            data = stations,
            weights = swipes2016)
#summary(ols1q)
```

Based on visual inspection, both the linear and quadratic models appear to fit the relationship between fare evasion arrest intensity and poverty rates across all stations fairly well. I prefer the quadratic model because it appears to fit the data slightly better, which is corroborated by a higher R-squared. The quadratic model has an adjusted R-squared of 0.21 compared to 0.15 for the linear model. The quadratic model also allows for curvature with informative results: arrest intensity is increasing in the poverty rate at an increasing rate (except for very low poverty rates).

Remember that we cannot interpret coefficients from a quadratic specification individually, we must jointly interpret and conduct a test of joint significance for the two poverty terms. The p-value for a test of joint significance is 1.2×10^{-4}, thus we can conclude that the quadratic terms are jointly significant.

If you prefer the linear specification because you find it simpler to interpret without changing the substantive conclusions, that is a reasonable justification. For the linear specification, it would make sense to interpret the slope coefficient on the poverty rate and do a t-test for statistical significance.

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

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

```
stations %>%
  ungroup() %>% #stations was already grouped by st_id, need to ungroup first
  group_by(highpov) %>%
  summarise(n = n(),
            mean_pov = weighted.mean(povrt_all_2016, swipes2016),
            mean_arrper = weighted.mean(arrperswipe, swipes2016))
```

```
## # A tibble: 2 x 4
##   highpov              n mean_pov mean_arrper
##   <fct>            <int>    <dbl>       <dbl>
## 1 Not high poverty    79    0.146       0.783
## 2 High poverty        77    0.319       1.42
```

```
#regress arrest intensity on highpov dummy to implement diff in means test
#weighted, robust SEs

ols_diff1 <- lm_robust(formula = arrperswipe ~ highpov,
                       data = stations,
                       weights = swipes2016)
#summary(ols_diff1)
```

The difference in average fare evasion arrest intensity between high- and low-poverty subway stations (weighted by ridership) is 0.63 with a p-value of 0.002. Thus we can conclude that this difference is statistically significant beyond the 1% level.

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 and 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. Present a similar table for the mean poverty rate by group in place of arrest intensity. Does it appear that differences in arrest intensity are explained by differences in poverty rate?

```
t1_arrper_wtd <- with(stations,
                      tapply(arrperswipe * swipes2016,
                             list("High Poverty" = highpov,
                                  "Predominantly Black" = nblack),
                             mean)/
                      tapply(swipes2016,
                             list("High Poverty" = highpov,
                                  "Predominantly Black" = nblack),
                             mean))

t1_povrt_wtd <- with(stations,
                  tapply(povrt_all_2016 * swipes2016,
                         list("High Poverty" = highpov,
                              "Predominantly Black" = nblack),
                         mean) /
                   tapply(swipes2016,
                         list("High Poverty" = highpov,
                              "Predominantly Black" = nblack),
                         mean))

round(t1_arrper_wtd, 2)
```

```
##                   Predominantly Black
## High Poverty       Majority non-Black Majority Black
##   Not high poverty               0.66           1.19
##   High poverty                   0.82           2.49
```

```
round(t1_povrt_wtd, 2)
```

```
##                   Predominantly Black
## High Poverty       Majority non-Black Majority Black
##   Not high poverty               0.13           0.19
##   High poverty                   0.32           0.32
```

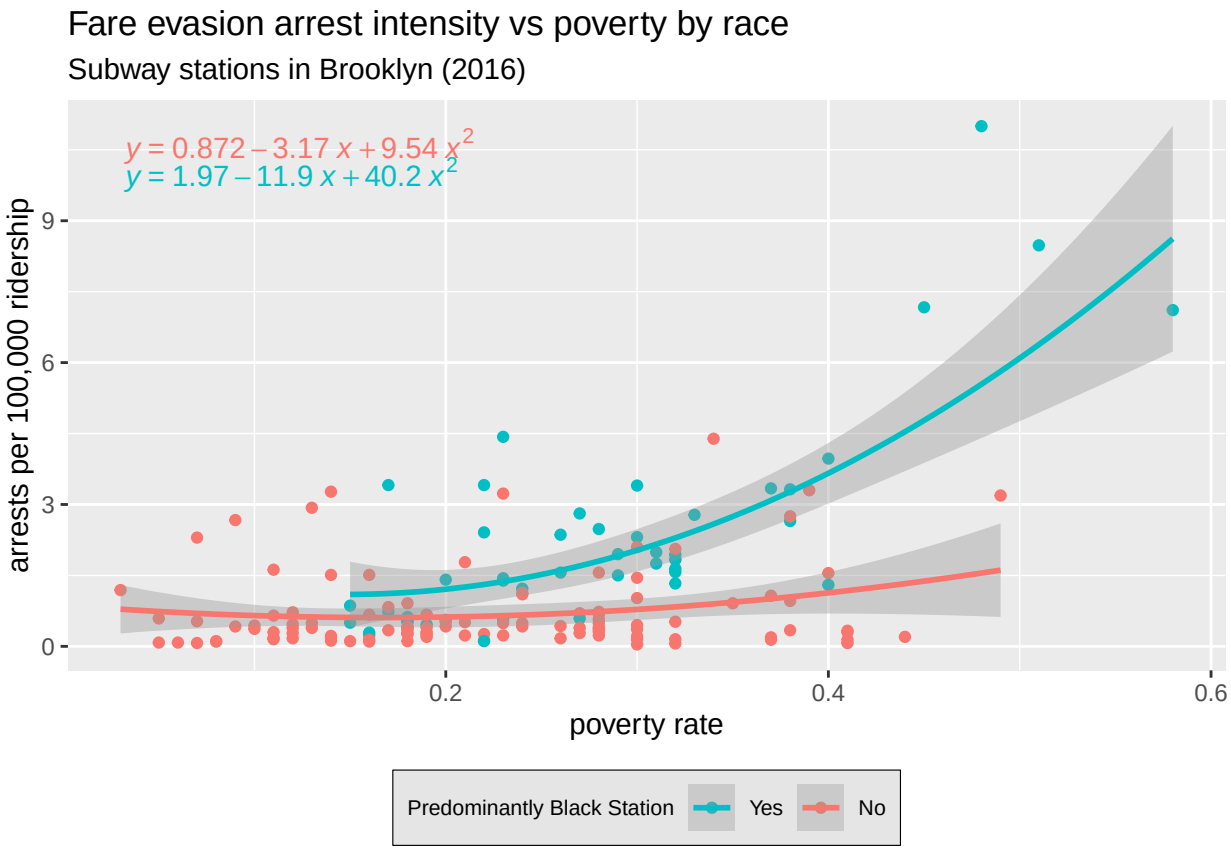
The above tables show that mean arrests per 100,000 ridership are more than 3 times as high at subway stations in majority Black areas compared to non-Black areas. Poverty rates, on the other hand, are very similar between majority-Black and non-Black high-poverty subway station areas, suggesting this is not a likely explanation for the difference in fare evasion arrest intensity (but we can use regression analysis to explore how the relationship between poverty rates and fare evasion differs based on neighborhood racial composition).

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

- Weight observations by ridership, and label your axes appropriately. **Only show one plot with your preferred specification!**
- Interpret your preferred regression specification (carefully!).

```
# specify quadratic formula to refer back to
  formula <-  y ~ poly(x, 2, raw = TRUE)

ggplot(stations, aes(x = povrt_all_2016,
                     y = arrperswipe,
                     weight = swipes2016,
                     color = nblack)) +
  geom_point()  +
  geom_smooth(method = 'lm',
              formula = formula) +  #add regression line
  stat_poly_eq(mapping = use_label("eq"),
               formula = formula)  +
  ylab("arrests per 100,000 ridership") +
  xlab("poverty rate") +
  ggtitle("Fare evasion arrest intensity vs poverty by race",
          subtitle = "Subway stations in Brooklyn (2016)") +
  scale_color_discrete(name   = "Predominantly Black Station",
                       labels = c("No", "Yes"),
                       guide  = guide_legend(reverse=TRUE)) +
  theme(legend.position = "bottom",
        legend.background = element_rect(color = "black",
                                         fill = "grey90",
                                         size = .2,
                                         linetype = "solid"),
          legend.direction = "horizontal",
          legend.text = element_text(size = 8),
          legend.title = element_text(size = 8))
```




```
# get separate data frames by predominantly Black stations to estimate separate models
  stations_black <- stations %>%
    filter(nblack == "Majority Black")

  stations_nonblack <- stations %>%
    filter(nblack == "Majority non-Black")

# nblack == 1: linear model with station observations
  ols_b_l <- lm_robust(arrperswipe ~ povrt_all_2016,
                       data = stations_black,
                       weights = swipes2016)

# nblack == 1: quadratic model with station observations
  ols_b_q <- lm_robust(arrperswipe ~ povrt_all_2016 + I(povrt_all_2016^2),
                       data = stations_black,
                       weights = swipes2016)

#nblack == 0: linear model with station observations
  ols_nb_l <- lm_robust(arrperswipe ~ povrt_all_2016,
                        data = stations_nonblack,
                        weights = swipes2016)

# nblack == 0: quadratic model with station observations
  ols_nb_q <- lm_robust(arrperswipe ~ povrt_all_2016 + I(povrt_all_2016^2),
                        se_type = "HC1",
                        data = stations_nonblack,
```

```
                        weights = swipes2016)
```

Visual inspection of the fitted regression lines reveal a clear pattern for both the linear and quadratic specifications: fare evasion arrest intensity increases (at an increasing rate) along with poverty rates at subway stations in predominantly Black areas, but not at other stations. Said another way, the result suggest that a predominantly Black station area tends to experience significantly higher arrest intensity than a non-Black station with a similarly high poverty rate.

Note that the above interpretation is qualitative in nature: it's a bit more straightforward to provide a numerical interpretation of coefficient estimates with a linear model. Alternatively, it would be informative to compare predicted fare evasion arrest intensity for a predominantly Black station area with a specified poverty rate (say, 40%) compared to a non-Black station area with the same poverty rate. If you prefer the linear specification because it is a bit simpler to interpret without changing the substantive conclusions, that is a reasonable justification.

I opt to present the quadratic specification ere; it explains 0.44) of the variation in fare evasion arrest intensity for predominantly Black station areas, compared to 0.51 for the linear specification. The p-value for a test of joint significance of the poverty coefficients in predominantly Black station areas is 1.8×10^{-4}, thus we can conclude that the poverty rate terms in the quadratic model are jointly significant. This compares to a p-value of 0.44381 for stations in non-Black station areas.

For both quadratic and linear models, poverty rates explain very little of the variation in arrest intensity among non-Black station areas in Brooklyn (0.02 and 0.01, respectively).

Regardless of functional form, poverty is only a statistically significant determinant of fare evasion arrest intensity at subway stations in predominantly Black station areas.

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?

We will discuss your answers and the issue of measurement error during class.

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. Drop the stations with the 4 highest crime counts, as they are in close proximity to the criminal courthouse and thus may experience higher arrest intensity for reasons unrelated to crime and poverty.

```
st_crime <- read.csv("nypd_criminalcomplaints_2016.csv")

stations_wcrime <- stations %>%
  inner_join(st_crime) %>%
  arrange(desc(crimes)) %>%
  filter(crimes < 2367) #exclude the stations with the 4 highest counts of criminal complaints
```

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.

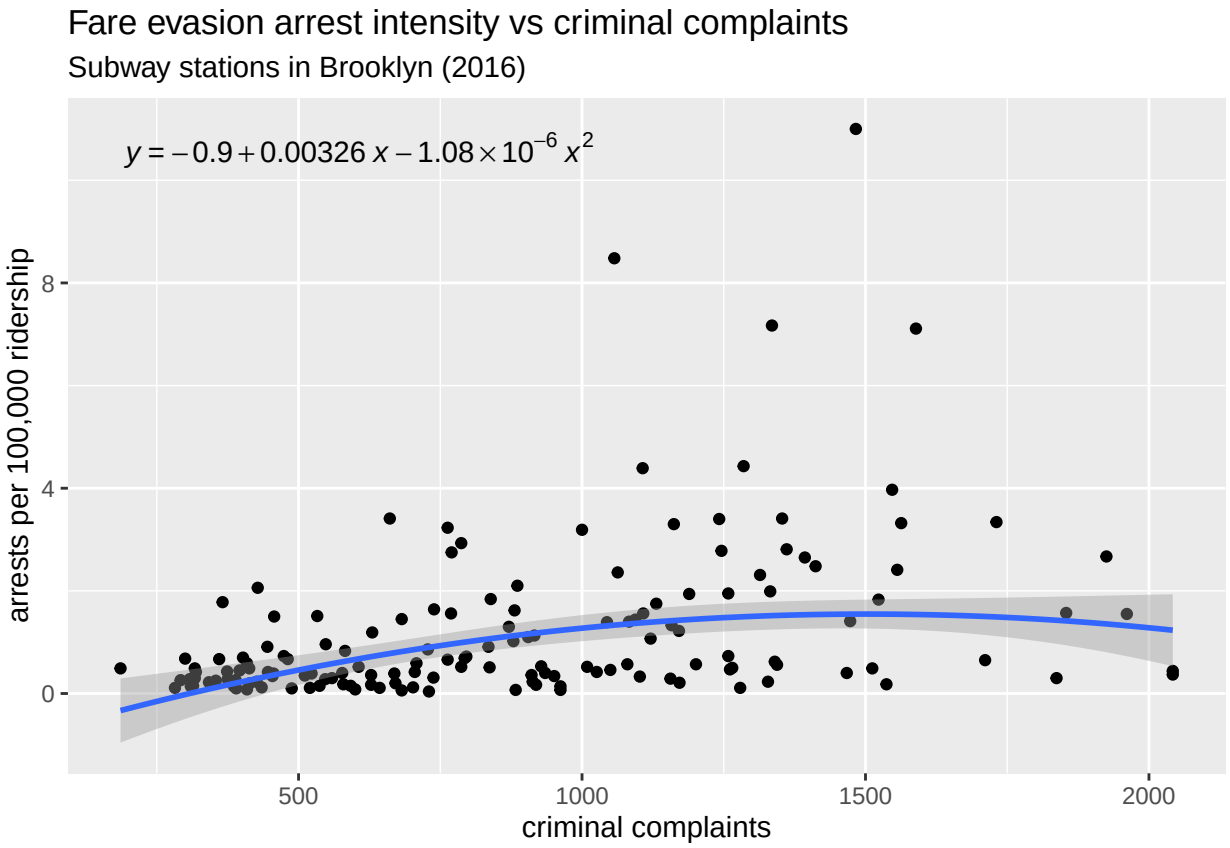
6b) First examine the overall relationship between arrest intensity and crime (without taking neighborhood racial composition or poverty into account) (comparable to Section 4a). Carefully interpret the results you choose to present.

- Show a scatterplot of your preferred crime measure vs. arrest intensity along with the regression line you think best fits this relationship. Weight observations by ridership, and label your axes appropriately. **Only show one plot with your preferred specification!**
- Interpret your preferred regression specification (carefully!). Remember to test for statistical significance for any estimates you choose to emphasize.

```
# specify quadratic formula to refer back to
  formula <-  y ~ poly(x, 2, raw = TRUE)

ggplot(stations_wcrime,
       aes(x = crimes,
           y = arrperswipe,
           weight = swipes2016)) +
  geom_point()  +
  geom_smooth(method = 'lm',
              formula = formula) +  #add regression line
  stat_poly_eq(mapping = use_label("eq"),
               formula = formula)  +
  ylab("arrests per 100,000 ridership") + xlab("criminal complaints") +
  ggtitle("Fare evasion arrest intensity vs criminal complaints",
          subtitle = "Subway stations in Brooklyn (2016)") +
  scale_color_discrete(name = "Predominantly Black Station",
                       labels=c("No", "Yes"),
                       guide = guide_legend(reverse=TRUE)) +
  theme(legend.position = "bottom",
        legend.background = element_rect(color = "black",
                                         fill = "grey90",
                                         size = .2,
                                         linetype = "solid"),
        legend.direction = "horizontal",
```

```
        legend.text = element_text(size = 8),
        legend.title = element_text(size = 8))
```




```
ols_c_l <- lm_robust(arrperswipe ~ crimes,
                     data = stations_wcrime,
                     weight = swipes2016)

ols_c_q <- lm_robust(arrperswipe ~ crimes + I(crimes^2),
                     data = stations_wcrime,
                     weight = swipes2016)
```

In the quadratic model, criminal complaints explain (0.133 of the variation in fare evasion arrest intensity across subway stations in Brooklyn.

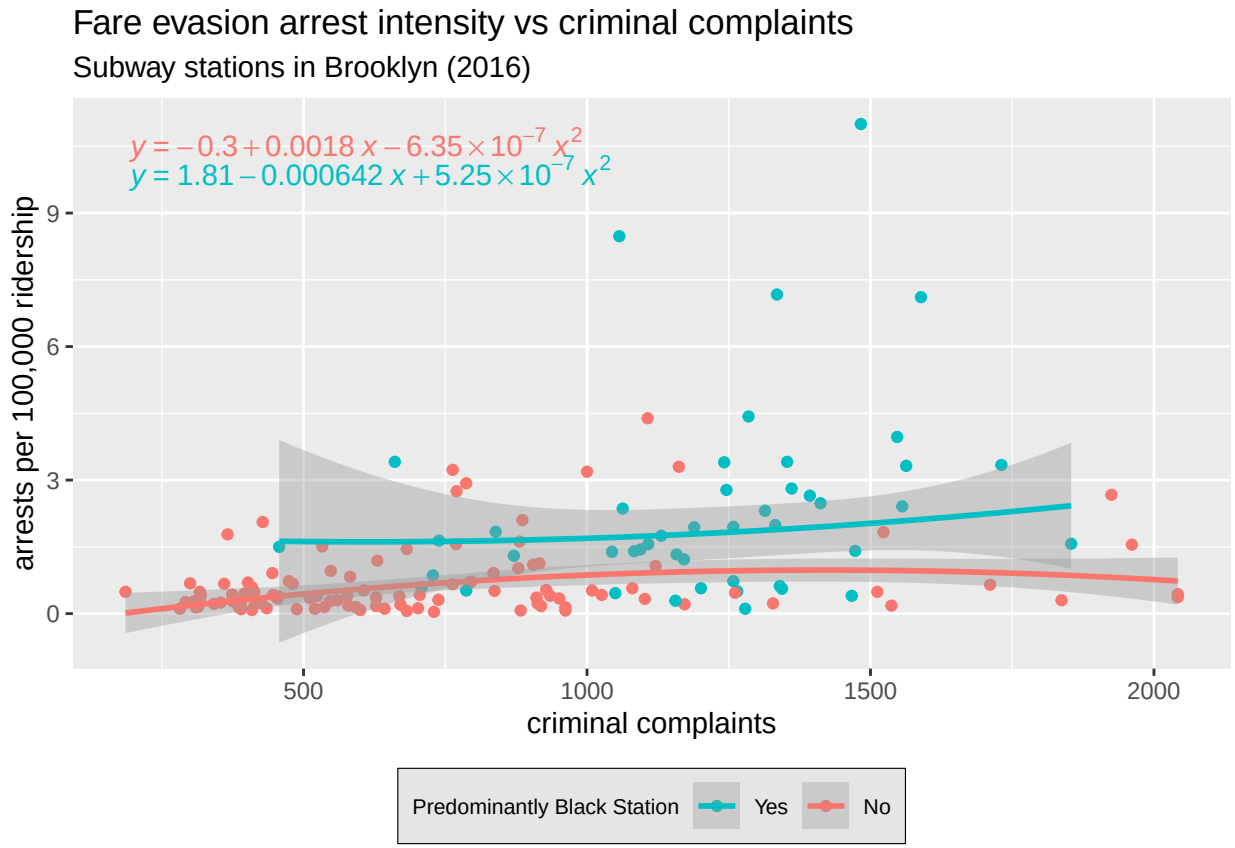
The p-value for a test of joint significance of the poverty coefficients is 0, thus we can conclude that the quadratic poverty rate terms are jointly significant.

6c) Examine how neighborhood racial composition mediates the relationship between arrest intensity and crime (comparable to Section 5b).

- Show a scatterplot of your preferred crime measure vs. arrest intensity along with the regression line you think best fits this relationship. Weight observations by ridership, and label your axes appropriately. **Only show one plot with your preferred specification!**
- Interpret your preferred regression specification (carefully!). Remember to test for statistical significance for any estimates you choose to emphasize.

```r
# specify quadratic formula to refer back to
  formula <-  y ~ poly(x, 2, raw = TRUE)

ggplot(stations_wcrime, aes(x = crimes,
                            y = arrperswipe,
                            weight = swipes2016,
                            color = nblack)) +
    geom_point()  +
  geom_smooth(method = 'lm',
              formula = formula) +  #add regression line
  stat_poly_eq(mapping = use_label("eq"),
               formula = formula)  +
    ylab("arrests per 100,000 ridership") + xlab("criminal complaints") +
    ggtitle("Fare evasion arrest intensity vs criminal complaints",
            subtitle = "Subway stations in Brooklyn (2016)") +
    scale_color_discrete(name = "Predominantly Black Station",
                         labels=c("No", "Yes"),
                         guide = guide_legend(reverse=TRUE)) +
    theme(legend.position = "bottom",
          legend.background = element_rect(color = "black", fill = "grey90",
                                           size = .2, linetype = "solid"),
          legend.direction = "horizontal",
          legend.text = element_text(size = 8),
          legend.title = element_text(size = 8))
```



```r
# get separate data frames by predominantly Black stations to estimate separate models
  stations_wcrime_black <- stations_wcrime %>%
    filter(nblack == "Majority Black")

  stations_wcrime_nonblack <- stations_wcrime %>%
    filter(nblack == "Majority non-Black")

# nblack == 1: linear model
ols_c_b_l <- lm_robust(arrperswipe ~ crimes,
                       data = stations_wcrime_black,
                       weight = swipes2016)

# nblack == 1: quadratic model
ols_c_b_q <- lm_robust(arrperswipe ~ crimes + I(crimes^2),
                       data = stations_wcrime_black,
                       weight = swipes2016)

# nblack == 0: linear model
ols_c_nb_l <- lm_robust(arrperswipe ~ crimes,
                        data = stations_wcrime_nonblack,
                        weight = swipes2016)

# nblack == 0: quadratic model
ols_c_nb_q <- lm_robust(arrperswipe ~ crimes + I(crimes^2),
                        data = stations_wcrime_nonblack,
                        weight = swipes2016)
```

Here we see that the prediction line is shifted up for predominantly Black station areas compared to other station areas, reflecting a greater predicted arrest intensity for every level of criminal complaints. However, the wider confidence bands for predominantly Black station areas indicate that the slope effects is imprecisely estimated. Indeed, the p-value for the joint significance of the crime terms in the quadratic model is 0.8308, so we cannot reject the null of no true effect of crime.

By comparison, for non-Black station areas the p-value for the joint significance of the crime terms in the quadratic model is 0.0369, indicating joint significance.

The key point here is that the predicted arrest intensity is higher in predominantly Black station areas for all levels of criminal complaints. This implies that differences in criminal complaints do not explain neighborhood racial disparities in arrest intensity.

Note that these results are particularly sensitive to the decision to weight observations by ridership. If you opted not to weight by ridership you will find that the slope effect of crime is indeed significant for predominantly Black station areas.

7 Summarize and interpret your findings with respect to subway fare evasion enforcement bias based on race. Be very careful about any claims of racial basis, any such claims should be supported by the analysis you present.

- 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.

The results presented here are consistent with race-based enforcement of fare evasion at subway stations in Brooklyn. As the poverty rate for a subway station area increases, fare evasion arrest intensity tends to increase in predominantly Black station areas (and the association is statistically significant) but not in non-Black station areas. For crime, the effect does not seem to be so pronounced, but neither do differences in crime explain differences in arrest intensity.

The analysis presented here does not support further conclusions about *why* poverty is effectively punished more intensively in predominantly Black station areas, though this result does not appear to be driven by differences in police presence to the extent we believe criminal complaints is a good proxy for police presence. Alternatively, the differential effect of poverty could be attributed to disparities in the decision to issue a summons rather than an arrest, perhaps due to explicit bias, implicit bias, or arrest quotas for NYPD Transit districts/sectors that correlate with neighborhood racial-ethnic composition. There may also be other differences in subway rider characteristics and behavior that could explain the observed relationship between neighborhood racial composition and fare evasion enforcement intensity, but disparate impact by race is apparent even if the all of the underlying mechanisms are not.

One additional test worth doing is to confirm that the positive association between poverty rates and fare evasion arrest intensity in predominantly Black neighborhoods is still statistically significant when simultaneously controlling for criminal complaints (but not in non-Black neighborhoods). This test confirms that, regardless of where the NYPD enforcement of other crimes is more prevalent, higher poverty Black neighborhoods face considerably higher fare evasion arrests than similarly higher poverty neighborhoods that are not predominantly Black.

Analyzing differences in fare evasion summonses compared to arrests would also be informative: are there significant differences in the demographics of individuals who are stopped for fare evasion, in addition to differences in the enforcement action taken once they are stopped? It would also be informative to see which communities are most affected during periods of time when the NYPD is "cracking down" more intensively on fare evasion.