

Data Science for Policy IA7514 Assignment 3

Subway Fare Evasion Arrests: Exploring Racial Disparities

Sample Solutions

2026-01-31

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

Load libraries

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

1 Aggregating to subway station-level arrest totals

1a) Load full set of cleaned arrest microdata (arrests.clean.rdata).

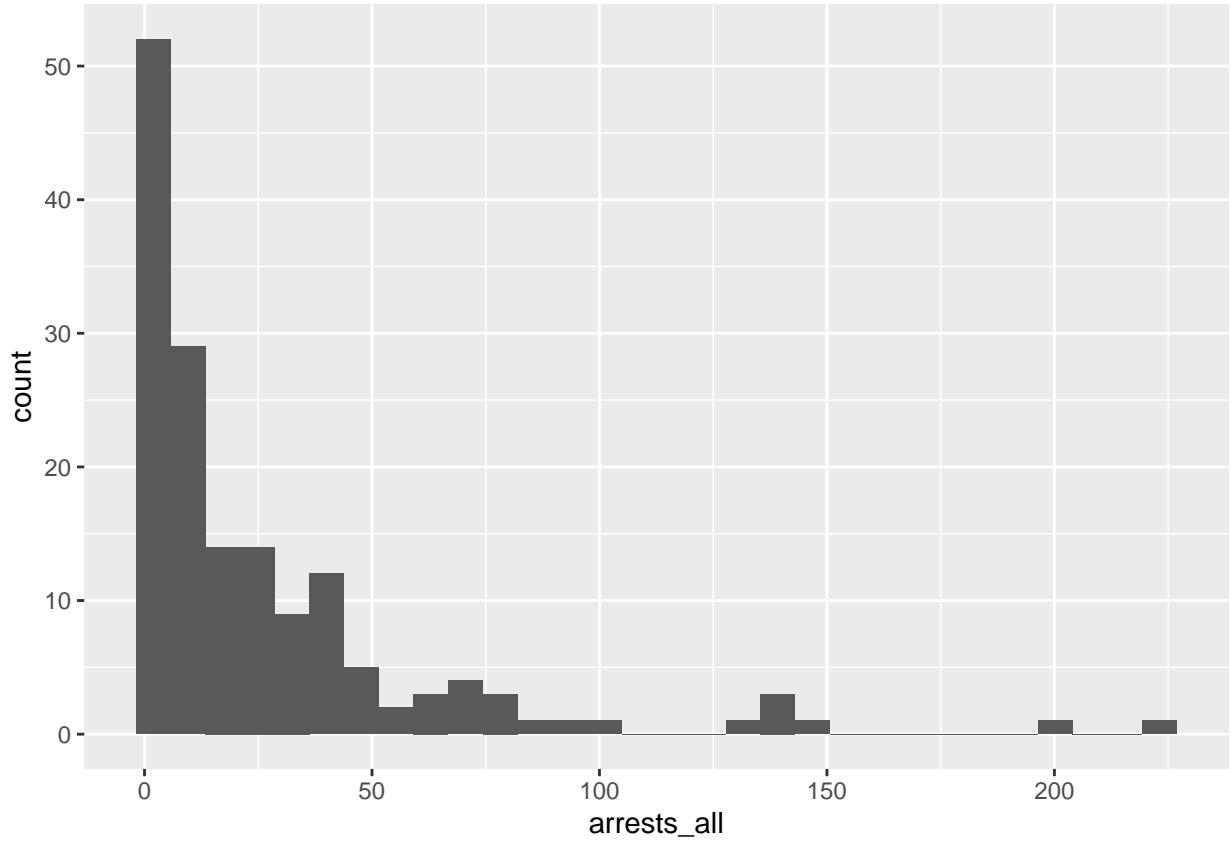
```
load("arrests.clean.RData")
```

1b) Using tidyverse functions, create a new data frame (st_arrests) that aggregates the micro-data 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))
```

1c) 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.

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

- 2a) 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)
```

- 2b) 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",...: 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 0 ...
## $ swipes2016   : int [1:157] 5025598 13091255 5152649 9051970 4272443 5861658 3897784 1435112 20317
```

2c) 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 better looking tables, we recommend passing your table into the `kable()` function from the `knitr` package. 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

2d) 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))
```

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

st_id	mta_name	arrperswip	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

2e) 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 highlights the importance of measuring arrests relative to ridership.

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

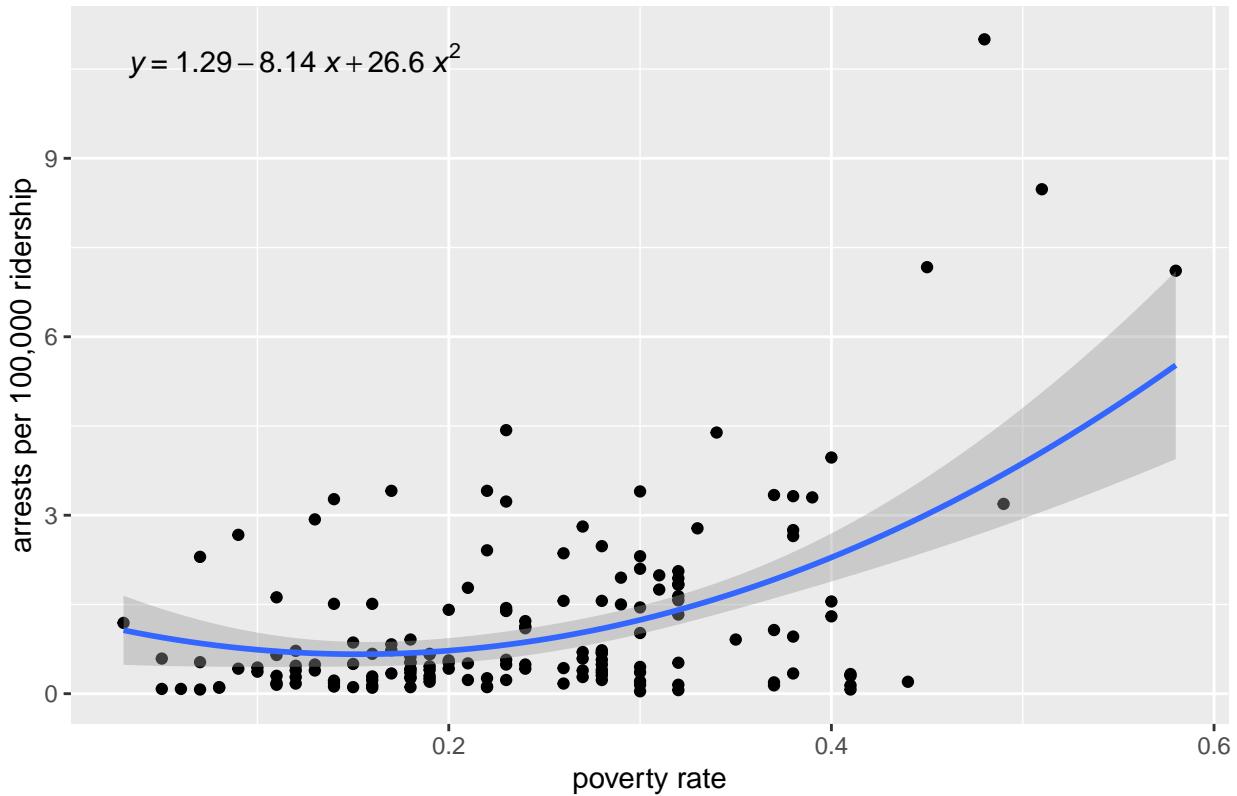
3a) Examine the relationship between arrest intensity and poverty rates

- Show a scatterplot of arrest intensity vs. poverty rates along with your preferred regression line (linear or quadratic, not both!). 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
prf <- 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 = prf) + #add regression line
stat_poly_eq(mapping = use_label("eq"),
             formula = prf)
```

Fare evasion arrest intensity vs. poverty rate



```
#linear model (all stations)
m_3a_l <- lm_robust(arrrperswipe ~ povrt_all_2016,
                      data = stations,
                      weights = swipes2016)

#quadratic model (all stations)
m_3a_q <- lm_robust(arrrperswipe ~ povrt_all_2016 + I(povrt_all_2016^2),
                      data = stations,
                      weights = swipes2016)
```

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 0.00012, 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, you could interpret the slope coefficient on the poverty rate and do a t-test for statistical significance.

3b) 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 = round(weighted.mean(povrt_all_2016, swipes2016), 2),
            mean_arrper = round(weighted.mean(arrperswipe, swipes2016), 2))

## # A tibble: 2 x 4
##   highpov           n  mean_pov  mean_arrper
##   <fct>     <int>    <dbl>      <dbl>
## 1 Not high poverty    79     0.15      0.78
## 2 High poverty        77     0.32      1.42

# regress arrest intensity on highpov dummy to implement diff in means test
# weight by ridership, estimate robust SEs
m_3b <- lm_robust(formula = arrperswipe ~ highpov,
                   data = stations,
                   weights = swipes2016)
```

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.

4 How does neighborhood racial composition (`nblack`) moderate the relationship between poverty and arrest intensity?

4a) Present a table showing the difference in mean arrests intensity for each of the four groups defined by the interaction of `highpov` and `nblack`. Remember to weight by ridership.

- HINT: use `group_by()` and `summarise()`
- BONUS: can you report this information in a 2x2 table?

```
stations %>%
  group_by(nblack, highpov) %>%
  summarise(mean = round(weighted.mean(arpperswipe, swipes2016), 2)) %>%
  pivot_wider(names_from = highpov,
              values_from = mean)

## # A tibble: 2 x 3
## # Groups:   nblack [2]
##   nblack           'Not high poverty' 'High poverty'
##   <fct>             <dbl>          <dbl>
## 1 Majority non-Black     0.66          0.82
## 2 Majority Black        1.19          2.49
```

4b) Does the difference in mean arrest intensity between high-poverty majority Black and high-poverty majority non-Black stations appear to be explained by differences in the mean poverty rate?

- **Step 1:** calculate the difference in mean arrest intensity between high poverty Majority Black and Majority Non-Black station areas. Make sure to calculate statistics weighted by ridership. Test whether this difference is statistically significant.
- **Step 2:** Repeat above steps for the poverty rate instead of arrest intensity.
- **Step 3:** Don't forget to answer the question above in your own words!

```
m_4b_arr <- lm_robust(formula = arpperswipe ~ nblack,
                        data = subset(stations,
                                      highpov == "High poverty"),
                        weights = swipes2016)
#summary(m_4b_arr)

m_4b_pov <- lm_robust(formula = povrt_all_2016 ~ nblack,
                        data = subset(stations,
                                      highpov == "High poverty"),
                        weights = swipes2016)
#summary(m_4b_pov)
```

The difference in the ridership-weighted mean arrest intensity between high-poverty majority Black station areas and high poverty majority non-Black station areas is 1.68 arrests per 100,000 ridership. This difference is statistically significant, with a p-value of 0.

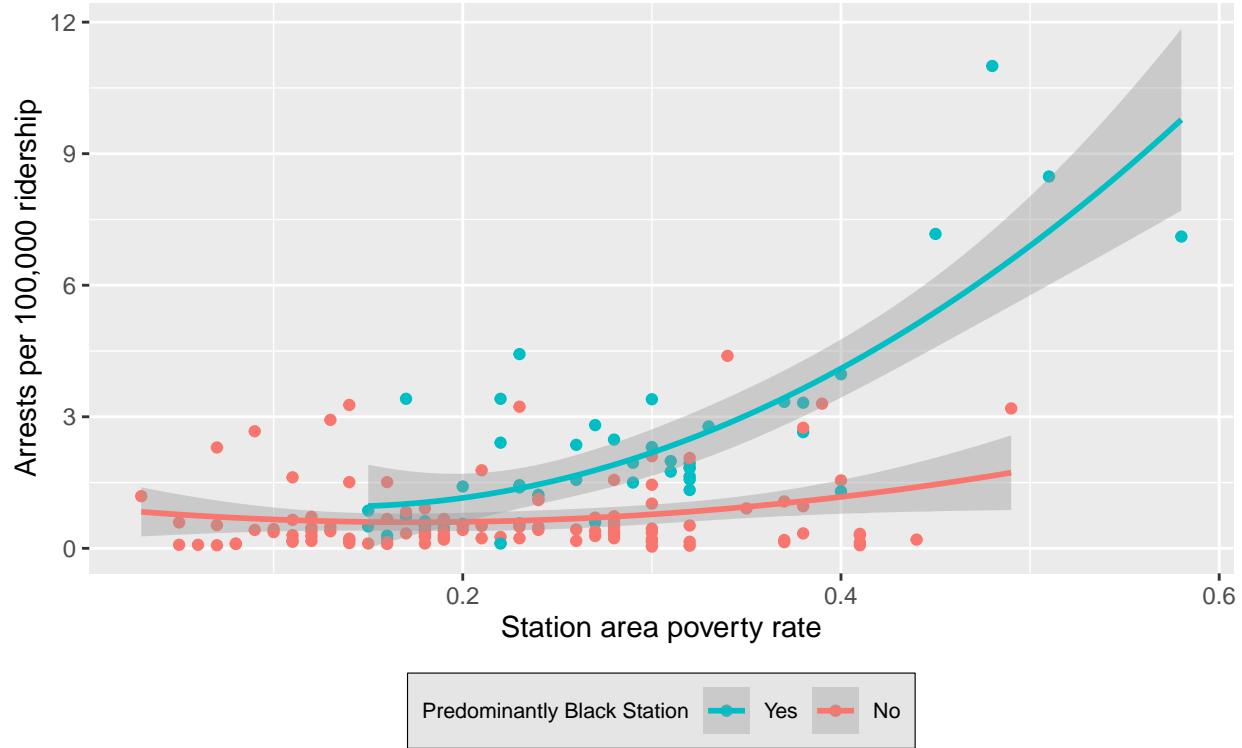
Could this difference in mean arrest intensity be explained by a higher mean poverty rate in high-poverty majority Black station areas? The difference in the ridership-weighted mean poverty rate between high-poverty majority Black station areas and high poverty majority non-Black station areas is only 0.004. This difference is not statistically significant, with a p-value of 0.7932. Thus it does not appear that the observed difference in mean arrest intensity is driven by a difference in the mean poverty rate.

4c) Present and interpret a scatterplot of arrest intensity vs. poverty rates broken down by majority Black vs. majority non-Black (`nblack`), including different regression lines for each group of stations.

- use separate aesthetics for Black and non-Black station areas
- include the regression lines that you think best capture this relationship:
- show linear or quadratic specifications (not both!)
- weight observations by ridership, and label your axes appropriately
- remember to carefully interpret your preferred regression specification (carefully!)

```
ggplot(stations,
       aes(x = povrt_all_2016,
           y = arrperswipe,
           color = nblack)) +
  geom_point() +
  geom_smooth(method = 'lm',
              formula = y ~ x + I(x^2)) +
  ylab("Arrests per 100,000 ridership") +
  xlab("Station area 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) )
```

Fare Evasion Arrest Intensity vs Poverty by Race Subway stations in Brooklyn (2016)



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

# nblack == 1: linear model with station observations
m_4c_black_l <- lm_robust(arpperswipe ~ povrt_all_2016,
                           data = stations_black,
                           weights = swipes2016)

# nblack == 1: quadratic model with station observations
m_4c_black_q <- lm_robust(arpperswipe ~ povrt_all_2016 + I(povrt_all_2016^2),
                           data = stations_black,
                           weights = swipes2016)

# nblack == 0: linear model with station observations
m_4c_nonblack_l <- lm_robust(arpperswipe ~ povrt_all_2016,
                               data = stations_nonblack,
                               weights = swipes2016)

# nblack == 0: quadratic model with station observations
m_4c_nonblack_q <- lm_robust(arpperswipe ~ povrt_all_2016 + I(povrt_all_2016^2),
                               data = stations_nonblack,
                               weights = swipes2016)

# or we can estimate an interaction model, interacting nblack and the pov rate
```

```
m_4c_interact <- lm_robust(arpperswipe ~ povrt_all_2016*nblack,
                             data = stations,
                             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 more straightforward to provide a numerical interpretation of coefficient estimates with a linear model. With the linear functional form, one can also estimate an interaction model to see that every 10 percentage point increase in the poverty rate is associated with an increase in arrests per 100,000 ridership that is 1.14 greater in majority Black station areas than non-Black station areas, and this difference in slope effects is statistically significant ($p\text{-value} = 0.00006$). Alternatively, one could compare the predicted fare evasion arrest intensity for a predominantly Black station area with a specified poverty rate (say, 40%) 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 here; it explains 0.51 percent of the variation in fare evasion arrest intensity for predominantly Black station areas, compared to 0.44 percent for the linear specification. The p -value for a test of joint significance of the poverty coefficients in predominantly Black station areas is 0.0002, thus we can conclude that the poverty rate terms in the quadratic model are jointly significant. This compares to a p -value of 0.4684 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. One interpretation of the observed associations is that poverty is effectively punished more intensively at the turnstile in majority Black station areas than in majority non-Black station areas.

4d) BONUS: Next let's think about how measurement error might impact results from 4c. Do you think measurement error could bias your estimates of neighborhood racial gaps in the effect of poverty on enforcement intensity from 4c? 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 its 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.

5 Is the differential effect of poverty in majority Black station areas explained by differences between stations in crime?

One determinant of fare evasion enforcement is police presence: when more police are present, the greater the chances they will encounter fare evasion. Moreover, the NYPD has often claimed they go where the crime is.

In the absence of data on police deployment across the subway system, we can use the number of crimes as a proxy for police presence.

5a) Load `nypd_crime_complaints_2016.csv` and join to stations by `st_id` and `mta_name`.

```
st_crime <- read.csv("nypd_crime_complaints_2016.csv")

stations <- inner_join(stations, st_crime) %>%
  group_by(st_id, mta_name) %>%
  arrange(desc(crimes))
```

5b) Are there more crimes reported in high-poverty Majority Black station areas than in high-poverty Majority non-Black station areas? Report the difference in crimes and assess statistically significance.

```
m_5b <- lm_robust(formula = crimes ~ nblack,
                    data = subset(stations,
                                  highpov == "High poverty"),
                    weights = swipes2016)
```

The average number of crimes in high-poverty, majority Black station areas is 496.1 greater than in high-poverty, majority non-Black station areas. This difference is statistically significant beyond the 1% level (p-value = 0.00007).

5c) Does the difference in crimes that you found in 5b explain the finding from 4c that poverty has a stronger positive effect on arrest intensity in majority Black station areas than in majority non-Black station areas?

- start with your preferred specification from 4c
- next, control for the the number of crimes and see if the conclusions from 4c change
- do your conclusions change if you consider different functional forms for the relationship between crime and arrest intensity?

```
# add the number of crimes as a linear control
m_5_interact_crime1 <- lm_robust(arrestswipe ~ povrt_all_2016*nblack + crimes,
                                    data = stations,
                                    weights = swipes2016)
summary(m_5_interact_crime1)

## 
## Call:
## lm_robust(formula = arrestswipe ~ povrt_all_2016 * nblack + crimes,
##           data = stations, weights = swipes2016)
##
```

```

## Weighted, Standard error type: HC2
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -0.1580137  0.2414414 -0.6545 0.513811
## povrt_all_2016              2.0543938  1.0219342  2.0103 0.046180
## nblackMajority Black      -1.6336707  0.7289923 -2.2410 0.026485
## crimes                     0.0004415  0.0001319  3.3468 0.001032
## povrt_all_2016:nblackMajority Black 9.9506151  2.7922959  3.5636 0.000490
##                                     CI Lower   CI Upper   DF
## (Intercept)                  -0.6350533  0.3190259 151
## povrt_all_2016                 0.0352573  4.0735303 151
## nblackMajority Black        -3.0740130 -0.1933284 151
## crimes                      0.0001809  0.0007022 151
## povrt_all_2016:nblackMajority Black 4.4335998 15.4676303 151
##
## Multiple R-squared:  0.4623 ,   Adjusted R-squared:  0.448
## F-statistic: 14.72 on 4 and 151 DF,  p-value: 3.543e-10
# add the log of crimes as a control
m_5_interact_crime2 <- lm_robust(arpperswipe ~ povrt_all_2016*nblack + log(crimes),
                                   data = stations,
                                   weights = swipes2016)
summary(m_5_interact_crime2)

##
## Call:
## lm_robust(formula = arpperswipe ~ povrt_all_2016 * nblack + log(crimes),
##           data = stations, weights = swipes2016)
##
## Weighted, Standard error type: HC2
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -2.9418    1.0451  -2.815 0.0055302
## povrt_all_2016               1.8603    1.0194   1.825 0.0700034
## nblackMajority Black      -1.7637    0.7416  -2.378 0.0186386
## log(crimes)                 0.4862    0.1537   3.164 0.0018793
## povrt_all_2016:nblackMajority Black 10.2053    2.7978   3.648 0.0003637
##                                     CI Lower   CI Upper   DF
## (Intercept)                  -5.0066  -0.8769 151
## povrt_all_2016                 -0.1539  3.8745 151
## nblackMajority Black        -3.2289 -0.2986 151
## log(crimes)                   0.1826  0.7898 151
## povrt_all_2016:nblackMajority Black 4.6775 15.7331 151
##
## Multiple R-squared:  0.4602 ,   Adjusted R-squared:  0.4459
## F-statistic: 16.06 on 4 and 151 DF,  p-value: 5.618e-11

```

No, differences in crime between stations do not explain the differential effect of the poverty rate on arrest intensity between majority Black and non-Black station areas. Even after controlling for the number of crimes (or the log of crimes), the positive slope effect of the poverty rate for majority Black station areas is significantly larger than for majority non-Black station areas. Said another way, differences in the incidence of crime do not seem to explain why poverty is effectively punished more intensively at the turnstile in majority Black station areas.

6. Summarize and interpret your findings with respect to racial disparities in subway fare evasion arrest intensity. Be very careful about how you frame and justify any claims of racial bias; any such claims should be supported by the analysis you present.

- Is there any additional analysis you'd like to do with the data at hand?
- Are there any key limitations to the data and/or analysis affecting your ability to examine racial disparities in enforcement?
- Is there any additional data you'd like to see that would help strengthen your analysis and interpretation?
- For this question, try to be very 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.