U6614: Assignment 4

Philip Crane (plc2137)

2024-02-17

Please submit your knitted .pdf file along with the corresponding R markdown (.rmd) via Courseworks by 11:59pm on Friday, February 17th.

1 Load libraries

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

getwd()
```

```
## [1] "C:/Users/philc/OneDrive/Desktop/Spring 2024/R/Assignments/Assignment4"
```

2 Aggregating to subway station-level arrest totals

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

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

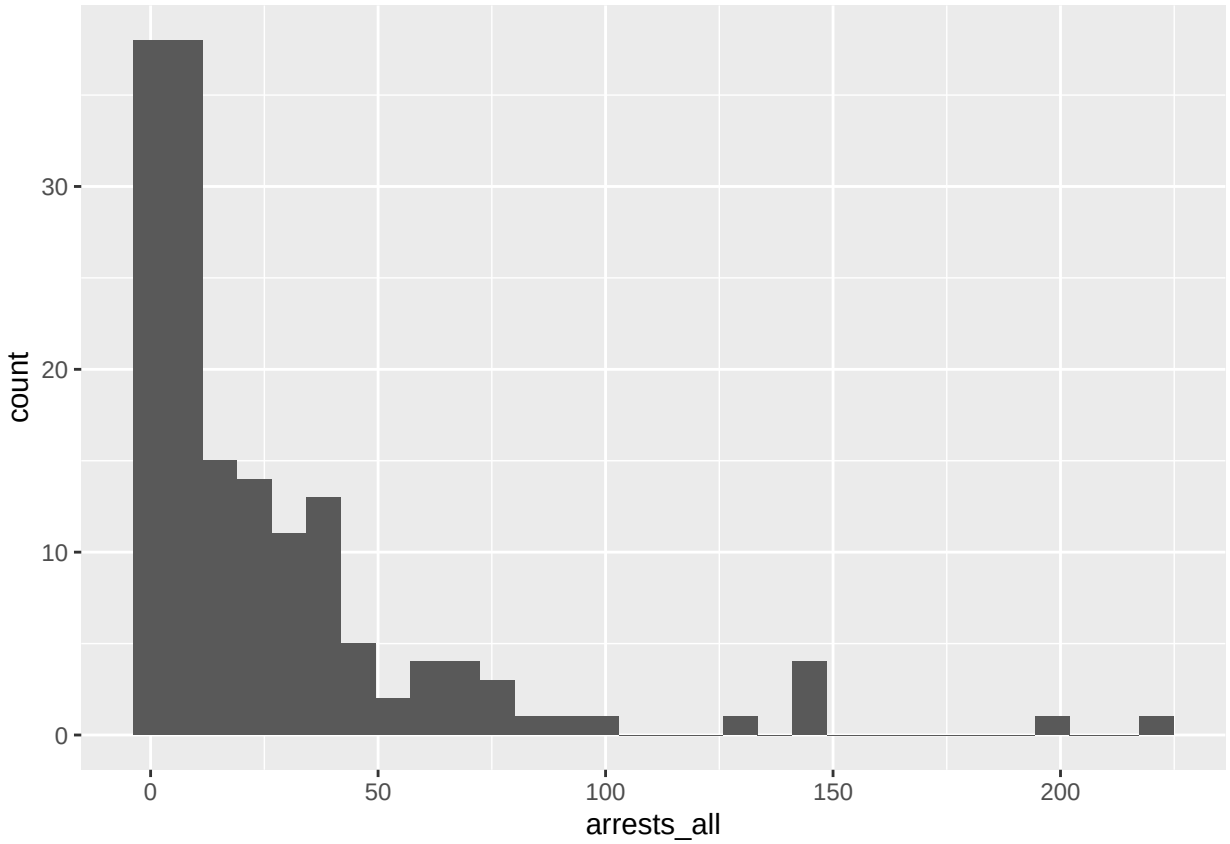
2b) Create new data frame (`st_arrests`) that aggregates microdata to station-level observations including the following information:

- `st_id`, `loc2`, 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(st_arrests, aes(x= arrests_all)) + geom_histogram()
```



The histogram shows an enormous difference between stations with the most arrests and those with the least. Coney Island has the most arrests by a good margin, and is the only station to exceed 200 arrests. The most frequent number of reported station arrests is 2.

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`).

```
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 columns 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_joinedtemp <- inner_join(st_arrests, st_poverty, by = c("st_id" = "st_id"))
  rm(st_joinedtemp)

st_joined <- st_arrests %>%
    inner_join(st_poverty, by = c("st_id" = "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)

st_joined %>% ungroup() %>% str(give.attr = FALSE)
```

```
## 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
```

```
    summary(st_joined)
```

```
##      st_id                                loc2      arrests_all
##  Min.   :  1   15 st prospect park f g line:  1   Min.   :  2.00
##  1st Qu.: 40   1523 Avenue U                :  1   1st Qu.:  4.00
##  Median : 79   1778 w 7th st                :  1   Median : 13.00
##  Mean   : 79   18th av and 85th st          :  1   Mean   : 26.82
##  3rd Qu.:118   18th ave and 64th st          :  1   3rd Qu.: 36.00
##  Max.   :157   20th ave and 64th st          :  1   Max.   :223.00
##                (Other)                      :151
##        x                y                                          mta_name
##  Min.   :-74.03   Min.   :40.58   15 St-Prospect Park F subway G subway:  1
##  1st Qu.:-73.98   1st Qu.:40.63   18 Av  D subway                      :  1
##  Median :-73.96   Median :40.67   18 Av  F subway                      :  1
##  Mean   :-73.96   Mean   :40.66   18 Av N subway                       :  1
##  3rd Qu.:-73.93   3rd Qu.:40.69   20 Av  D subway                      :  1
##  Max.   :-73.87   Max.   :40.72   20 Av  N subway                      :  1
##                                   (Other)                             :151
##  pop_black_2016  pov_black_2016  pop_all_2016    pov_all_2016
##  Min.   :    0   Min.   :   0   Min.   : 2721   Min.   :  308
##  1st Qu.:  398   1st Qu.:  86   1st Qu.:11303   1st Qu.: 1959
##  Median : 2399   Median : 616   Median :15167   Median : 3206
##  Mean   : 4579   Mean   :1166   Mean   :15250   Mean   : 3775
##  3rd Qu.: 7504   3rd Qu.:1644   3rd Qu.:18371   3rd Qu.: 5241
##  Max.   :20739   Max.   :6706   Max.   :31071   Max.   :11612
##
##  povrt_all_2016      shareblack          nblack        swipes2016
##  Min.   :0.03321   Min.   :0.00000   Min.   :0.000   Min.   :  406793
##  1st Qu.:0.16348   1st Qu.:0.02366   1st Qu.:0.000   1st Qu.: 1188884
##  Median :0.22859   Median :0.15819   Median :0.000   Median : 1863036
##  Mean   :0.24010   Mean   :0.29328   Mean   :0.293   Mean   : 2449301
##  3rd Qu.:0.30144   3rd Qu.:0.54165   3rd Qu.:1.000   3rd Qu.: 3027658
##  Max.   :0.57800   Max.   :0.88898   Max.   :1.000   Max.   :13818168
##
```

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)

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

```
## # A tibble: 10 x 5
## # Groups:   st_id, mta_name [10]
##    st_id mta_name                          arrests_all shareblack povrt_all_2016
##    <int> <fct>                                   <int>      <dbl>          <dbl>
##  1    66 "Coney Island-Stillwell Av D sub~         223    0.00694          0.256
##  2    99 "Jay St-MetroTech A subway C sub~         198    0.156            0.156
##  3   150 "Utica Av  A subway C subway "            143    0.799            0.331
##  4    70 "Crown Heights-Utica Av 3 subway~         142    0.748            0.317
```

```
##  5   114 "Marcy Av  J subway M subway Z s~        141    0.0650          0.387
##  6   131 "Nostrand Av  A subway C subway"        141    0.647           0.220
##  7    54 "Canarsie-Rockaway Pkwy L subway"       133    0.833           0.171
##  8   147 "Sutter Av  L subway"                   102    0.749           0.578
##  9   106 "Kingston-Throop Avs C subway"           90    0.785           0.233
## 10   123 "Nevins St  2 subway 3 subway 4 ~        86    0.198           0.0894
```

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

```
stations <- st_joined %>%
    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),
           shareblack = round(shareblack, 2),
           povrt_all_2016 = round(povrt_all_2016, 2)) %>%
    mutate(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"))) %>%
    filter(st_id != 66)

stations %>%
    arrange(desc(arrperswipe)) %>%
    select(st_id, mta_name, arrperswipe, arrests_all, shareblack, povrt_all_2016, highpov, nblack) %>%
    head(n = 10) %>%
    kable()
```

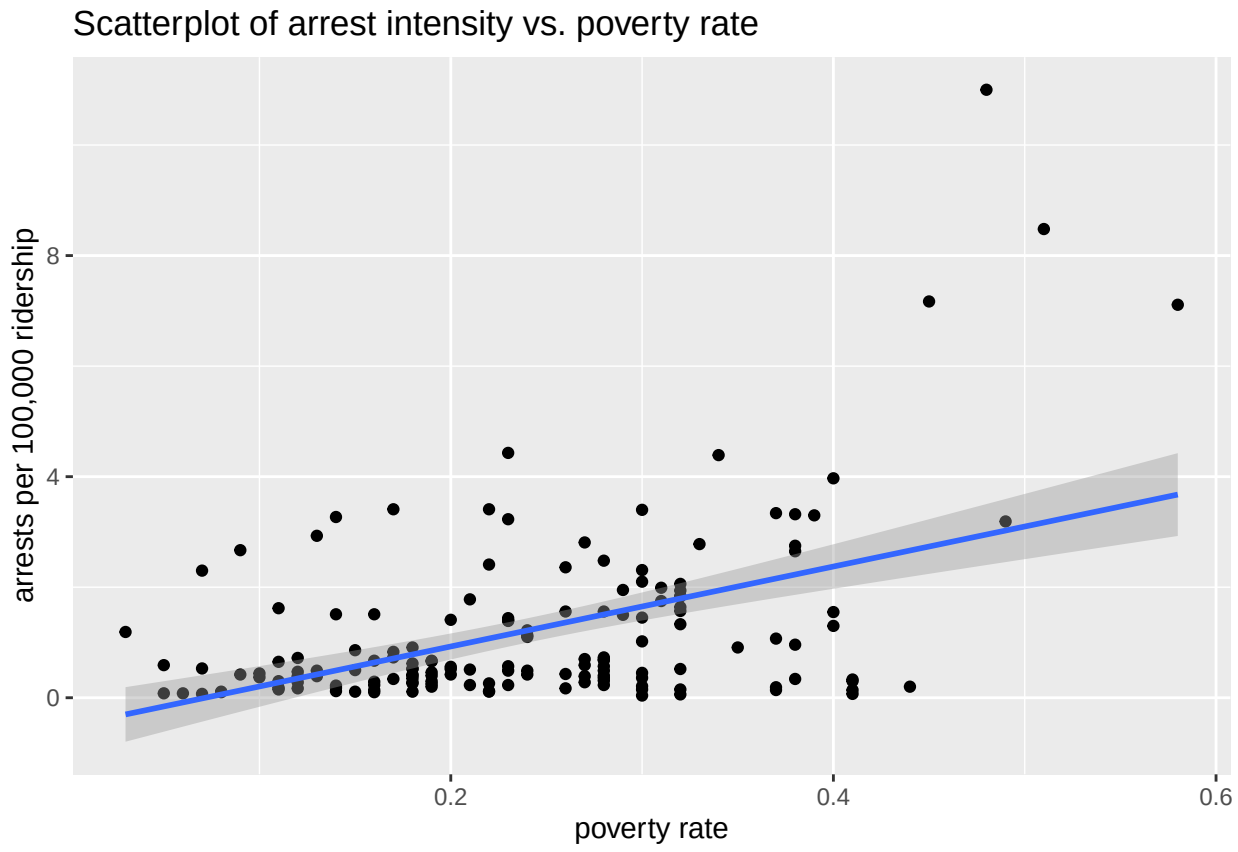
st_id	mta_name	arrperswipe	arrests_all	shareblack	povrt_all_2016	highpov	nblack
101	Junius St 3 subway	11.00	75	0.78	0.48	High poverty	Majority Black
26	Atlantic Av L subway	8.48	37	0.66	0.51	High poverty	Majority Black
111	Livonia Av L subway	7.17	75	0.83	0.45	High poverty	Majority Black
147	Sutter 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

st_id	mta_name	arrperswipe	arrests_all	shareblack	povrt_all_2016	highpov	nblack
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

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!).

```
ggplot(stations,
       aes(x = povrt_all_2016, y = arrperswipe)) +
  geom_point() +
  ggtitle('Scatterplot of arrest intensity vs. poverty rate') +
  labs(x = 'poverty rate', y = 'arrests per 100,000 ridership') +
  geom_smooth(method = 'lm', formula=y~x)
```



```
  #geom_smooth(method = "lm", formula = y~x + I(x^2))
```

```
ols1l <- lm(arrperswipe ~ povrt_all_2016, data = stations, weights = swipes2016)
ols1q <- lm(arrperswipe ~ povrt_all_2016 + I(povrt_all_2016^2),
            data = stations, weights = swipes2016)

summary(ols1l)
```

```
##
## Call:
## lm(formula = arrperswipe ~ povrt_all_2016, data = stations, weights = swipes2016)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -3197.5  -904.3  -364.7   541.3  7222.3
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.04023    0.21273   0.189     0.85
## povrt_all_2016  4.61213    0.88237   5.227 5.53e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1731 on 154 degrees of freedom
## Multiple R-squared:  0.1507, Adjusted R-squared:  0.1452
## F-statistic: 27.32 on 1 and 154 DF,  p-value: 5.528e-07
```

```
coeftest(ols1l, vcov = vcovHC(ols1l, type = "HC1"))
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    0.040233   0.242262  0.1661 0.8683166
## povrt_all_2016 4.612128   1.182665  3.8998 0.0001434 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(ols1q)
```

```
##
## Call:
## lm(formula = arrperswipe ~ povrt_all_2016 + I(povrt_all_2016^2),
##     data = stations, weights = swipes2016)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -4044.5  -863.9  -330.4   656.8  6182.4
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)           1.2865     0.3841   3.349 0.001020 **
## povrt_all_2016       -8.1441     3.4376  -2.369 0.019080 *
## I(povrt_all_2016^2)  26.6321     6.9564   3.828 0.000188 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1659 on 153 degrees of freedom
## Multiple R-squared:  0.2249, Adjusted R-squared:  0.2148
## F-statistic:  22.2 on 2 and 153 DF,  p-value: 3.426e-09
```

```
coeftest(ols1q, vcov = vcovHC(ols1q, type = "HC1"))
```

```
##
## t test of coefficients:
##
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.28648    0.39608  3.2480 0.001428 **
## povrt_all_2016       -8.14412    3.99163 -2.0403 0.043040 *
## I(povrt_all_2016^2) 26.63207    9.07709  2.9340 0.003862 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this case, the linear specification seems to better fit the model because the non-linear regression suggests a negative correlation (-8.14412) between poverty and arrests per swipe. This over-complicates relationship we want to show in this particular model, while the linear model shows a far more straightforward positive relationship between the two variables. Conversely, the quadratic R-squared of 0.2249 is higher than the linear R-squared of 0.1507, suggesting higher goodness of fit in the quadratic regression. I opted to add weights to "swipes2016" as significant measurement error could be occurring in that variable (i.e. riders in heavily-trafficked stations not being reflective of the poverty of the surrounding neighborhood).

Regression interpretation: On average, a 1% increase in poverty is associated with a 4.612 increase in arrests per 100,000 swipes. This is statistically significant at the 99% confidence interval.

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

```
diff1 <- lm(arrperswipe ~ highpov, data = stations, weight = swipes2016)
    summary(diff1)
```

```
##
## Call:
## lm(formula = arrperswipe ~ highpov, data = stations, weights = swipes2016)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -2947.1  -916.2  -410.8   656.2  7914.3
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.7829     0.1225   6.392 1.85e-09 ***
```

```
## highpovHigh poverty   0.6332     0.1884   3.360 0.000981 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1813 on 154 degrees of freedom
## Multiple R-squared:  0.06832,    Adjusted R-squared:  0.06227
## F-statistic: 11.29 on 1 and 154 DF,  p-value: 0.0009815
```

```
    coeftest(diff1, vcov = vcovHC(diff1, type="HC1"))
```

```
##
## t test of coefficients:
##
##                     Estimate Std. Error t value  Pr(>|t|)
## (Intercept)          0.78293    0.11127  7.0365 6.077e-11 ***
## highpovHigh poverty  0.63316    0.19953  3.1732  0.001821 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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?

```
t1_arrper_wtd <-
  with(stations,
       tapply(arrperswipe * swipes2016,
              list(highpov, nblack),
              sum))  /
  with(stations,
       tapply(swipes2016,
              list(highpov, nblack),
              sum) )

t1_arrper_wtd <- t1_arrper_wtd %>% round(2)
t1_arrper_wtd
```

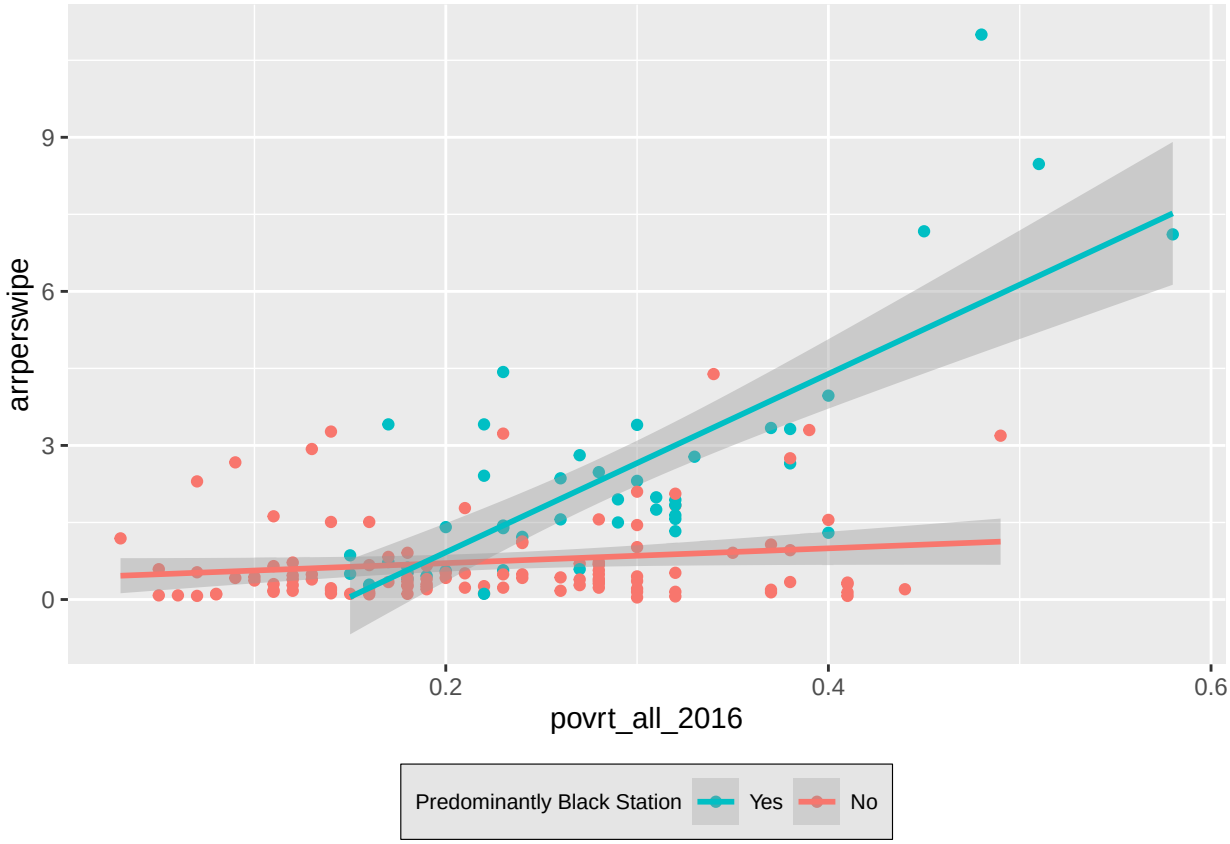
```
##                  Majority non-Black Majority Black
## Not high poverty               0.66           1.19
## High poverty                   0.82           2.49
```

Race alone and poverty alone does not explain differences in arrest rates. As shown above, the interaction of race and poverty accounts for the biggest differences in arrest rates.

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!).

```
ggplot(stations, aes(x = povrt_all_2016, y = arrperswipe, color = nblack)) +
    geom_point() +
    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 = "soli
          legend.direction = "horizontal",
          legend.text = element_text(size = 8),
          legend.title = element_text(size = 8) ) +
    geom_smooth(method = 'lm', formula=y~x)
```




```
    #geom_smooth(method = "lm", formula = y~x + I(x^2))
```

```
stations_black <- stations %>%
  filter(nblack == "Majority Black")

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

ols2qb <- lm(arrperswipe ~ povrt_all_2016 + I(povrt_all_2016^2), data = stations_black)
ols2qnb <- lm(arrperswipe ~ povrt_all_2016 + I(povrt_all_2016^2), data = stations_non_black)

ols2lb <- lm(arrperswipe ~ povrt_all_2016, data = stations_black)
ols2lnb <- lm(arrperswipe ~ povrt_all_2016, data = stations_non_black)

summary(ols2qb)
```

```
##
## Call:
## lm(formula = arrperswipe ~ povrt_all_2016 + I(povrt_all_2016^2),
##     data = stations_black)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8004 -0.6455 -0.2596  0.2268  4.7304
##
```

```
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.731      1.644   1.053  0.29813
## povrt_all_2016       -11.737     10.560  -1.112  0.27253
## I(povrt_all_2016^2)   44.150     15.713   2.810  0.00743 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.346 on 43 degrees of freedom
## Multiple R-squared:  0.6493, Adjusted R-squared:  0.633
## F-statistic: 39.81 on 2 and 43 DF,  p-value: 1.642e-10
```

```
coeftest(ols2qb, vcov = vcovHC(ols2qb, type = "HC1"))
```

```
##
## t test of coefficients:
##
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.7312     1.7502  0.9891  0.32814
## povrt_all_2016      -11.7371    12.7088 -0.9235  0.36088
## I(povrt_all_2016^2)  44.1503    22.2482  1.9844  0.05361 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(ols2qnb)
```

```
##
## Call:
## lm(formula = arrperswipe ~ povrt_all_2016 + I(povrt_all_2016^2),
##     data = stations_non_black)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1954 -0.4587 -0.2425  0.0692  3.4834
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.9399     0.3716   2.530   0.0129 *
## povrt_all_2016       -3.9510     3.3919  -1.165   0.2467
## I(povrt_all_2016^2)  11.3320     6.9337   1.634   0.1051
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8433 on 107 degrees of freedom
## Multiple R-squared:  0.05345,    Adjusted R-squared:  0.03576
## F-statistic: 3.021 on 2 and 107 DF,  p-value: 0.05293
```

```
coeftest(ols2qnb, vcov = vcovHC(ols2qnb, type = "HC1"))
```

```
##
## t test of coefficients:
##
```

```
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.93994    0.37169  2.5288   0.0129 *
## povrt_all_2016      -3.95105    3.58582 -1.1019   0.2730
## I(povrt_all_2016^2) 11.33200    8.02686  1.4118   0.1609
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(ols2lb)
```

```
##
## Call:
## lm(formula = arrperswipe ~ povrt_all_2016, data = stations_black)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0929 -0.8625 -0.3157  0.4367  5.2176
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2.5547     0.6589  -3.877 0.000349 ***
## povrt_all_2016  17.3689     2.2057   7.875 6.11e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.448 on 44 degrees of freedom
## Multiple R-squared:  0.5849, Adjusted R-squared:  0.5755
## F-statistic: 62.01 on 1 and 44 DF,  p-value: 6.111e-10
```

```
coeftest(ols2lb, vcov = vcovHC(ols2lb, type = "HC1"))
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    -2.55469    0.80562 -3.1711  0.002766 **
## povrt_all_2016 17.36890    3.17360  5.4729 1.995e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(ols2lnb)
```

```
##
## Call:
## lm(formula = arrperswipe ~ povrt_all_2016, data = stations_non_black)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9411 -0.4738 -0.2912  0.0913  3.4798
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.4200     0.1935   2.171   0.0321 *
```

```
## povrt_all_2016   1.4418     0.7913   1.822   0.0712 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8498 on 108 degrees of freedom
## Multiple R-squared:  0.02982,    Adjusted R-squared:  0.02084
## F-statistic:  3.32 on 1 and 108 DF,  p-value: 0.07123
```

```
coeftest(ols2lnb, vcov = vcovHC(ols2lnb, type = "HC1"))
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.42002    0.21097  1.9909  0.04901 *
## povrt_all_2016  1.44176    0.98866  1.4583  0.14766
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Again the linear model seems to have better explanatory power particularly for majority-Black stations as the t-tests have more statistical significance. Both the intercept and povrt_all_2016 are statistically significant at the .001 and 0 levels for majority-Black stations, while the quadratic counterpart is not statistically significant. The R-quared is still higher in the quadratic regressions at 0.6493 and 0.05345 versus the linear R-squared values of 0.5849 and 0.02982. With negigible differences in the statistical significance of non-Black station coefficients across the board, one might suggest that while the linear model does a better job explaining majority-black stations, the quadratic regression has more explanaory power for non-majority black stations.

Interpretation majority-Black: On average, a 1% increase in poverty in majority-black areas is associated with a 17.36890 increase in arrests per 100,000 swipes. This is statistically significant at the 0 level.

Interpretation non-Black: On average, a 1% increase in poverty in majority non-Black areas is associated with a 1.44176 increase in arrests per 100,000 swipes. This is not statistically significant.

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?

One major source of measurement error could be rider transience, where heavily-trafficked stations are not necessarily composed of riders originating from the surrounding neighborhoods, therefore not reflective of

the surrounding poverty. Earlier, we discussed that the majority of those arrested are under 30. This could point to another instance of measurement error where those arrested who are under-18 may be processed via juvenile courts, therefore not appearing in this data. A third point of measurement error is that those arrested in more affluent neighborhoods may be less likely to use public defender services.

One way of combating this may be to collect socioeconomic data on the arrestees themselves rather than relying purely on the poverty indexes of the areas travelers happen to be arrested in. Another mechanism could be to look at overpolicing by controlling for how many police officers are present in any particular station.

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.

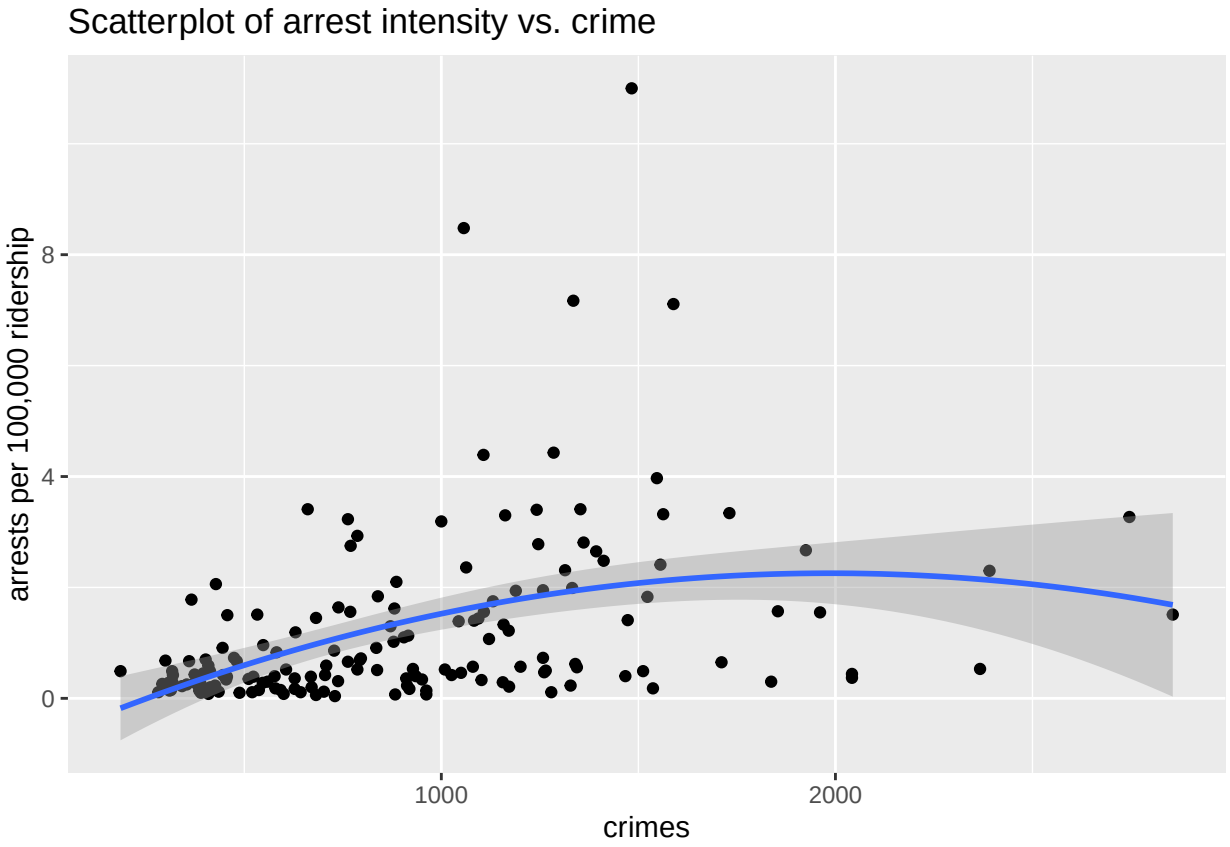
```
st_complaint <- read.csv("nypd_criminalcomplaints_2016.csv",
                         stringsAsFactors = TRUE)

stations <-
    inner_join(stations, st_complaint, by = c("st_id" = "st_id"))
```

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.

```
ggplot(stations,
       aes(x = crimes, y = arrperswipe)) +
  geom_point() +
  ggtitle('Scatterplot of arrest intensity vs. crime') +
  labs(x = 'crimes', y = 'arrests per 100,000 ridership') +
  #geom_smooth(method = 'lm', formula=y~x)
  geom_smooth(method = "lm", formula = y~x + I(x^2))
```




```
#ols3l <- lm(arrperswipe~crimes, data=stations, weights=swipes2016)

#summary(ols3l)
#coeftest(ols3l, vcov = vcovHC(ols2l, type="HC1"))

ols3q <- lm(arrperswipe ~ crimes + I(crimes^2),
            data = stations, weights = swipes2016)
summary(ols3q)
```

```
##
## Call:
## lm(formula = arrperswipe ~ crimes + I(crimes^2), data = stations,
##     weights = swipes2016)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -4939.4  -795.9   -65.2   845.5  7831.9
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.778e-01  3.483e-01  -1.085 0.279792
## crimes       2.051e-03  5.609e-04   3.656 0.000352 ***
## I(crimes^2) -5.217e-07  1.883e-07  -2.771 0.006289 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1762 on 153 degrees of freedom
## Multiple R-squared:  0.1263, Adjusted R-squared:  0.1149
## F-statistic: 11.06 on 2 and 153 DF,  p-value: 3.274e-05
```

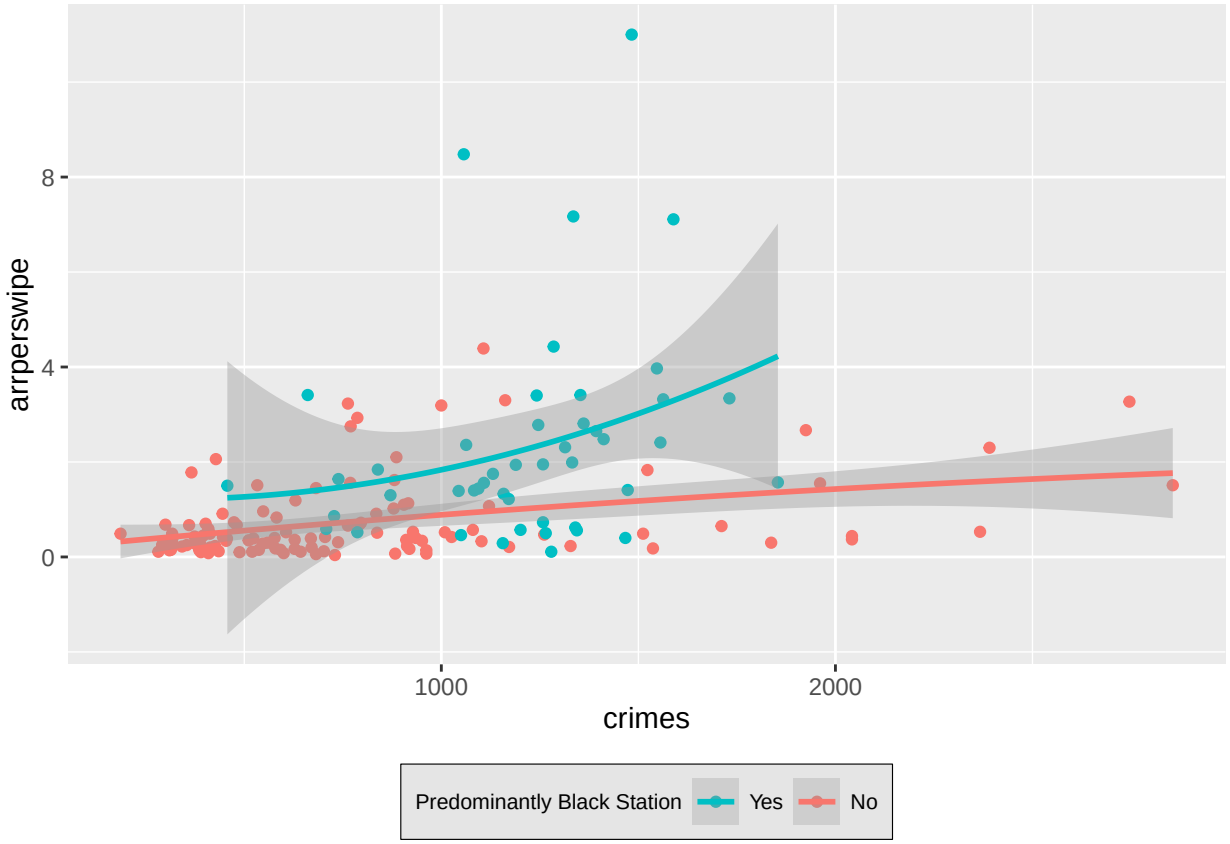
```
coeftest(ols3q, vcov = vcovHC(ols3q, type="HC1"))
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) -3.7779e-01  2.9300e-01 -1.2894 0.1992091
## crimes       2.0506e-03  5.6675e-04  3.6181 0.0004026 ***
## I(crimes^2) -5.2165e-07  1.8427e-07 -2.8308 0.0052668 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation: On average, for every additional crime reported, there is an associated 2.05 increase in arrests per 100,000 swipes. This is statistically significant at the 0 level.

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.

```
ggplot(stations, aes(x = crimes, y = arrperswipe, color = nblack)) +
    geom_point() +
    #Modify legend title and text
    scale_color_discrete(name = "Predominantly Black Station",
                         labels=c("No", "Yes"),
                         #Reverse Label Order
                         guide = guide_legend(reverse=TRUE)) +
    #Modify legend aesthetics (optional)
    theme(legend.position = "bottom",
          legend.background = element_rect(color = "black", fill = "grey90", size = .2, linetype = "soli
          legend.direction = "horizontal",
          legend.text = element_text(size = 8),
          legend.title = element_text(size = 8) ) +
    geom_smooth(method = "lm", formula = y~x + I(x^2))
```




```
ols4l <- lm(arrperswipe ~ crimes + nblack + (nblack * crimes), data = stations)
  summary(ols4l)
```

```
##
## Call:
## lm(formula = arrperswipe ~ crimes + nblack + (nblack * crimes),
##     data = stations)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5037 -0.5457 -0.2700  0.2141  8.0623
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   0.2747539  0.2323348   1.183   0.2388
## crimes                        0.0005770  0.0002402   2.402   0.0175 *
## nblackMajority Black         -0.4870588  0.8912327  -0.547   0.5855
## crimes:nblackMajority Black   0.0015471  0.0007334   2.109   0.0365 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.345 on 152 degrees of freedom
## Multiple R-squared:  0.2892, Adjusted R-squared:  0.2751
## F-statistic: 20.61 on 3 and 152 DF,  p-value: 2.931e-11
```

```r
coeftest(ols4l, vcov = vcovHC(ols4l, type = "HC1"))
```

```
##
## t test of coefficients:
##
##                                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)                  0.27475389  0.11564853  2.3758 0.0187589 *
## crimes                       0.00057698  0.00015757  3.6616 0.0003453 ***
## nblackMajority Black        -0.48705879  1.10292270 -0.4416 0.6594009
## crimes:nblackMajority Black  0.00154711  0.00100483  1.5397 0.1257189
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation: On average, each additional crime reported is associated with a 0.00057698 increase in arrests per 100,000 swipes. Majority-Black stations are associated -0.48705879 decrease in arrests per 100,000 swipes, while the interaction term of crime and Majority-Black stations indicates that for every additional crime reported near a majority-Black station, there is a 0.00154711 increase in arrests per 100,000 swipes.

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.

Overwhelmingly, race and poverty seem to have the most explanatory power in predicting arrest intensity. While a positive correlation does exist between crime reports and fare evaision arrests, the relationship is far weaker and does not have the statistical significance of the previous regressions. Possible additional analyisis with this data would be to keep the "swipes2011", "swipes2012", "swipes2013", "swipes2014", "swipes2015" columns to look at how ridership has grown across all stations against the 2016 arrest data. This could highlight how increased ridership in some stations might result in more police presence, therefore more arrests. Another line of analysis could be to create an interaction term between "crimes" and "povrt_all_2016" to run a regression similar to that in 6c.

An important data piece that is missing from all this is number of police deployed to particular stations. This would allow us to see how stations are being policed, and if overpoliced stations coincide with those in high-poverty areas, majority-Black areas, or areas with high reported crime. Another much-needed piece of information is the amount of citations and warnings issued in instances of fare evasion, in addition to arrests. One might hypothesize that including warnings and citations could illustrate systemic differences in how minority vs. non-minority fare evaders are treated, and if police in certain stations are more prone to arresting rather than issuing a citation.