School test scores and local poverty levels

Sara Ho

2020-09-22

1. Import data and packages

Import necessary packages.

```
library(tidyverse)
library(data.table)
```

Load the data

```
schools <- fread(here::here("../data", "nys_schools.csv"))
counties <- fread(here::here("../data", "nys_acs.csv"))
```

2. Explore and clean data

```
summary(schools)
```

```
##    school_cd            school_name        district_name
##  Min.   : 10100010014   Length:35663       Length:35663
##  1st Qu.:280210030004   Class :character   Class :character
##  Median :331700011533   Mode  :character   Mode  :character
##  Mean   :356790938901
##  3rd Qu.:472506040001
##  Max.   :680801040002
##  county_name           region               year       total_enroll
##  Length:35663       Length:35663       Min.   :2008   Min.   : -99.0
##  Class :character   Class :character   1st Qu.:2010   1st Qu.: 339.0
##  Mode  :character   Mode  :character   Median :2013   Median : 469.0
##                                        Mean   :2013   Mean   : 523.6
##                                        3rd Qu.:2015   3rd Qu.: 648.0
##                                        Max.   :2017   Max.   :2347.0
##  per_free_lunch     per_reduced_lunch      per_lep          mean_ela_score
##  Min.   :-99.0000   Min.   :-99.00000   Min.   :-99.00000   Min.   :-99.0
##  1st Qu.:  0.1900   1st Qu.:  0.03000   1st Qu.:  0.00000   1st Qu.:296.0
##  Median :  0.4200   Median :  0.06000   Median :  0.03000   Median :324.2
##  Mean   :  0.4188   Mean   :  0.02852   Mean   :  0.04124   Mean   :447.1
##  3rd Qu.:  0.7200   3rd Qu.:  0.10000   3rd Qu.:  0.11000   3rd Qu.:666.3
##  Max.   :257.0000   Max.   : 53.00000   Max.   :  1.00000   Max.   :720.8
##  mean_math_score
```

```
##  Min.   :-99.0
##  1st Qu.:298.0
##  Median :330.8
##  Mean   :456.0
##  3rd Qu.:683.5
##  Max.   :738.7
```

Deal with missing values, which are currently coded as `-99`.

```
# since we need county level data later, let's remove schools without a county
schools <- schools[county_name != -99]

# for remaining columns, convert -99 to NA
metric_names <- c("total_enroll", "per_free_lunch", "per_reduced_lunch", "per_lep", "mean_ela_score", "
schools[, (metric_names) := replace(.SD, .SD == -99, NA)
        , .SDcols = metric_names]

# verify here that there are no -99 in the minimum values
summary(schools)
```

```
##    school_cd           school_name        district_name
##  Min.   : 10100010014   Length:35644       Length:35644
##  1st Qu.:280210030004   Class :character   Class :character
##  Median :331700011533   Mode  :character   Mode  :character
##  Mean   :356870982249
##  3rd Qu.:472506040001
##  Max.   :680801040002
##
##  county_name           region               year       total_enroll
##  Length:35644       Length:35644       Min.   :2008   Min.   :   3.0
##  Class :character   Class :character   1st Qu.:2010   1st Qu.: 339.0
##  Mode  :character   Mode  :character   Median :2013   Median : 469.0
##                                        Mean   :2013   Mean   : 523.9
##                                        3rd Qu.:2015   3rd Qu.: 648.0
##                                        Max.   :2017   Max.   :2347.0
##                                                       NA's   :6
##  per_free_lunch     per_reduced_lunch     per_lep        mean_ela_score
##  Min.   :  0.0000   Min.   : 0.00000   Min.   :0.00000   Min.   :191.0
##  1st Qu.:  0.1900   1st Qu.: 0.03000   1st Qu.:0.00000   1st Qu.:300.0
##  Median :  0.4200   Median : 0.06000   Median :0.03000   Median :347.7
##  Mean   :  0.4605   Mean   : 0.07018   Mean   :0.07735   Mean   :483.2
##  3rd Qu.:  0.7200   3rd Qu.: 0.10000   3rd Qu.:0.11000   3rd Qu.:667.3
##  Max.   :257.0000   Max.   :53.00000   Max.   :1.00000   Max.   :720.8
##  NA's   :8          NA's   :8          NA's   :6         NA's   :2196
##  mean_math_score
##  Min.   :213.0
##  1st Qu.:303.0
##  Median :361.7
##  Mean   :492.7
##  3rd Qu.:684.7
##  Max.   :738.7
##  NA's   :2198
```

Since these columns cannot be over 100%, convert numbers above 1 to fractions. If there are remaining columns still over 1, replace these data points with `NA`.

```
schools[per_free_lunch > 1, per_free_lunch := per_free_lunch / 100]
schools[per_free_lunch > 1, per_free_lunch := NA]

schools[per_reduced_lunch > 1, per_reduced_lunch := per_reduced_lunch / 100]
schools[per_reduced_lunch > 1, per_reduced_lunch := NA]

# verify these variables are no larger than 1
summary(schools[, per_reduced_lunch])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.00000 0.03000 0.06000 0.06865 0.10000 0.93000       8
```

```
summary(schools[, per_free_lunch])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.1900  0.4200  0.4513  0.7200  1.0000       9
```

Recode county poverty variable

Create a categorical variable `pov_cat` that groups counties into "high", "medium", and "low" poverty groups. For each year, designate the 25th percentile as the cutoff for low poverty, and the 75th percentile as the cutoff for high poverty.

First, obtain the low and high cutoffs for each year.

```
# `poverty_cutoffs` contains the low and high cutoffs for each year in `counties`
poverty_cutoffs <- counties[ , .(cutoff_low = quantile(county_per_poverty, 0.25),
                                 cutoff_high = quantile(county_per_poverty, 0.75)), year]
poverty_cutoffs
```

```
##    year cutoff_low cutoff_high
## 1: 2009  0.1057346   0.1424311
## 2: 2010  0.1068775   0.1399779
## 3: 2011  0.1062462   0.1480074
## 4: 2012  0.1085408   0.1494247
## 5: 2013  0.1112546   0.1507370
## 6: 2014  0.1177611   0.1560124
## 7: 2015  0.1195301   0.1550593
## 8: 2016  0.1167323   0.1512235
```

Then use the cutoffs to create `pov_cat`.

```
# merge `poverty_cutoffs` with `counties`
counties <- counties[poverty_cutoffs, on = "year"]

# create the `pov_cat` variable
counties[, pov_cat := "medium"]
counties[county_per_poverty < cutoff_low, pov_cat := "low"]
```

```
counties[county_per_poverty > cutoff_high, pov_cat := "high"]

# delete the extra cutoff variables
counties[, cutoff_low := NULL][, cutoff_high := NULL]
head(counties)
```

```
##    county_name year county_per_poverty median_household_income county_per_bach
## 1:      ALBANY 2009          0.1183511                    55350      0.19036819
## 2:    ALLEGANY 2009          0.1521532                    40917      0.09468291
## 3:       BRONX 2009          0.2710533                    33794      0.11091251
## 4:      BROOME 2009          0.1427803                    43467      0.14127256
## 5: CATTARAUGUS 2009          0.1506553                    41482      0.09627803
## 6:      CAYUGA 2009          0.1148711                    47414      0.11163666
##    pov_cat
## 1:  medium
## 2:    high
## 3:    high
## 4:    high
## 5:    high
## 6:  medium
```

Create helpful additional variables in `school`

The tests that the NYS Department of Education administers changes from time to time, so scores are not directly comparable year-to-year. Create a new variable that is the standardized z-score for math and English Language Arts (ELA) for each year.

```
schools[, z_mean_ela_score := scale(mean_ela_score), year]
schools[, z_mean_math_score := scale(mean_math_score), year]
```

Create variables `num_free_lunch` and `num_reduced_lunch` that represent the total number of students in free and reduced lunch.

```
schools[ , num_free_lunch := round(total_enroll * per_free_lunch)]
schools[ , num_reduced_lunch := round(total_enroll * per_reduced_lunch)]
```

Also, create a variable `per_free_reduced_lunch` that sums the percentages of students in free lunch programs and students in reduced lunch programs

```
# create `per_free_reduced_lunch`
schools[, per_free_reduced_lunch := per_free_lunch + per_reduced_lunch]

# For some schools, `per_free_reduced_lunch` is greater than 1.
head(schools[per_free_reduced_lunch > 1, .(per_free_reduced_lunch, per_free_lunch, per_reduced_lunch)])
```

```
##    per_free_reduced_lunch per_free_lunch per_reduced_lunch
## 1:                   1.01           0.91              0.10
## 2:                   1.02           0.77              0.25
## 3:                   1.04           0.78              0.26
## 4:                   1.18           0.87              0.31
## 5:                   1.03           0.97              0.06
## 6:                   1.04           0.92              0.12
```

```
# it's possible that students in the reduced lunch category here are included in the "free lunch" categ
# so let's replace `per_free_reduced_lunch` with `per_free_lunch` for these rows.
schools[per_free_reduced_lunch > 1, per_free_reduced_lunch := per_free_lunch]
summary(schools$per_free_reduced_lunch)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.2500  0.5200  0.5198  0.8000  1.0000       9
```

Merge datasets

Create a county-level data set that merges variables from the schools dataset and the ACS dataset.

```
merged <- merge(schools, counties, by = c("county_name", "year"), all.x = TRUE)
```

---

3. Create summary tables

For each county: total enrollment, percent of students qualifying for free or reduced price lunch, and percent of population in poverty.

First, get the number of total students per county enrolled in free or reduced lunch program:

```
# mean county poverty across the years for which there is data
sum_table_county <- merged[ , .(tot_enroll = sum(total_enroll, na.rm = TRUE),
                            tot_reduced_lunch = sum(num_reduced_lunch, na.rm = TRUE),
                            tot_free_lunch = sum(num_free_lunch, na.rm = TRUE),
                            mean_poverty = mean(county_per_poverty, na.rm = TRUE)), county_name]
```

Then, convert the total numbers to percentages (fractions)

```
sum_table_county[, p_reduced_lunch := tot_reduced_lunch / tot_enroll]
sum_table_county[, p_free_lunch := tot_free_lunch / tot_enroll]
sum_table_county[, tot_reduced_lunch := NULL][, tot_free_lunch := NULL]
head(sum_table_county)
```

```
##    county_name tot_enroll mean_poverty p_reduced_lunch p_free_lunch
## 1:      ALBANY     257192    0.1229838      0.05341924    0.3344272
## 2:    ALLEGANY      55903    0.1508549      0.11180080    0.4062573
## 3:       BRONX    1646130    0.2872294      0.05528603    0.8032233
## 4:      BROOME     193424    0.1584756      0.07619013    0.3987716
## 5: CATTARAUGUS      93698    0.1642522      0.10651241    0.3966147
## 6:      CAYUGA      58462    0.1148119      0.07064418    0.3541446
```

For the counties with the top 5 and bottom 5 poverty rate: percent of population in poverty, percent of students qualifying for free or reduced price lunch,mean reading score, and mean math score.

We'll do this for the most current year with county data, which is 2016

First, select the counties with the top 5 and bottom 5 poverty rates from 2016, the most recent year.

```
# slice the county dataset into `low_pov_counties` and `high_pov_counties`, each contain 5 counties
low_pov_counties <- counties[year == max(year)][order(county_per_poverty)] %>% slice(1:5)
high_pov_counties <- counties[year == max(year)][order(county_per_poverty, decreasing = TRUE)] %>% slic
```

Scores are not comparable from year to year, but they should be comparable from school to school as long as the years are the same!

Check the data

Low poverty schools

```
low_pov_schools <- schools[low_pov_counties, on = c("county_name", "year")]
summary(low_pov_schools$mean_ela_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   271.2   302.0   312.3   310.8   320.1   346.0      67
```

```
summary(low_pov_schools$mean_math_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   258.2   300.7   315.3   313.0   324.6   355.0      67
```

```
n_distinct(low_pov_schools$school_cd)
```

```
## [1] 642
```

High poverty schools

```
high_pov_schools <- schools[high_pov_counties, on = c("county_name", "year")]
summary(high_pov_schools$mean_ela_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   259.3   289.7   297.7   299.9   308.7   351.3      38
```

```
summary(high_pov_schools$mean_math_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   247.7   282.3   294.0   297.0   309.0   375.3      38
```

```
n_distinct(high_pov_schools$school_cd)
```

```
## [1] 831
```

From this information, we know that:

- From the bottom 5 low poverty counties, there are 642 schools, of which 67 are missing ELA and math scores.
- From the top 5 high poverty counties, there are 831 schools, of which 38 are missing ELA and math scores.

It looks like missing score data is not correlated with high poverty levels.

Create a summary table for the schools from low poverty counties and a summary table for the schools from high poverty counties.

```
# create low poverty summary table
sum_table_low_pov <- low_pov_schools[ , .(tot_enroll = sum(total_enroll, na.rm = TRUE),
                     tot_reduced_lunch = sum(num_reduced_lunch, na.rm = TRUE),
                     tot_free_lunch = sum(num_free_lunch, na.rm = TRUE),
                     mean_poverty = mean(county_per_poverty, na.rm = TRUE),
                     mean_math_score = mean(mean_math_score, na.rm = TRUE),
                     mean_ela_score = mean(mean_ela_score, na.rm = TRUE))
                  , county_name][, pov_cat := "low"]

# create high poverty summary table
sum_table_high_pov <- high_pov_schools[ , .(tot_enroll = sum(total_enroll, na.rm = TRUE),
                      tot_reduced_lunch = sum(num_reduced_lunch, na.rm = TRUE),
                      tot_free_lunch = sum(num_free_lunch, na.rm = TRUE),
                      mean_poverty = mean(county_per_poverty, na.rm = TRUE),
                      mean_math_score = mean(mean_math_score, na.rm = TRUE),
                      mean_ela_score = mean(mean_ela_score, na.rm = TRUE))
                   , county_name][, pov_cat := "high"]

# convert totals to percentages
sum_table_low_pov[, p_free_reduced_lunch := (tot_reduced_lunch + tot_free_lunch)/ tot_enroll]
sum_table_high_pov[, p_free_reduced_lunch := (tot_reduced_lunch + tot_free_lunch)/ tot_enroll]

# remove the columns with total student numbers
sum_table_low_pov[, tot_enroll := NULL][, tot_reduced_lunch := NULL][, tot_free_lunch := NULL]
sum_table_high_pov[, tot_enroll := NULL][, tot_reduced_lunch := NULL][, tot_free_lunch := NULL]
```

Here are the resulting summary tables:

```
sum_table_low_pov
```

```
##    county_name mean_poverty mean_math_score mean_ela_score pov_cat
## 1:      PUTNAM   0.05091140        313.3393       312.0060     low
## 2:      NASSAU   0.05917006        320.1810       316.8058     low
## 3:    SARATOGA   0.06309991        318.3476       312.2846     low
## 4:     SUFFOLK   0.07108796        307.6212       306.5373     low
## 5:    DUTCHESS   0.08638583        301.7847       302.8058     low
##    p_free_reduced_lunch
## 1:            0.2077812
## 2:            0.2885859
## 3:            0.2403204
## 4:            0.3852053
## 5:            0.3295711
```

```
sum_table_high_pov
```

```
##    county_name mean_poverty mean_math_score mean_ela_score pov_cat
## 1:       BRONX    0.2976889        291.0076       295.7927    high
## 2:       KINGS    0.2253745        300.7693       303.6666    high
```

```
## 3:  MONTGOMERY    0.2016228        293.2167       289.5083    high
## 4:  CHAUTAUQUA    0.1852780        300.4387       297.4485    high
## 5:      OSWEGO    0.1750848        302.3482       294.3958    high
##    p_free_reduced_lunch
## 1:            0.8422316
## 2:            0.7201768
## 3:            0.5712829
## 4:            0.5709186
## 5:            0.5684146
```

4.1 Data visualization and analysis (lunch programs)

> What can the data tell us about the relationship between poverty and test performance in New York public schools? Has this relationship changed over time? Is this relationship at all moderated by access to free/reduced price lunch?

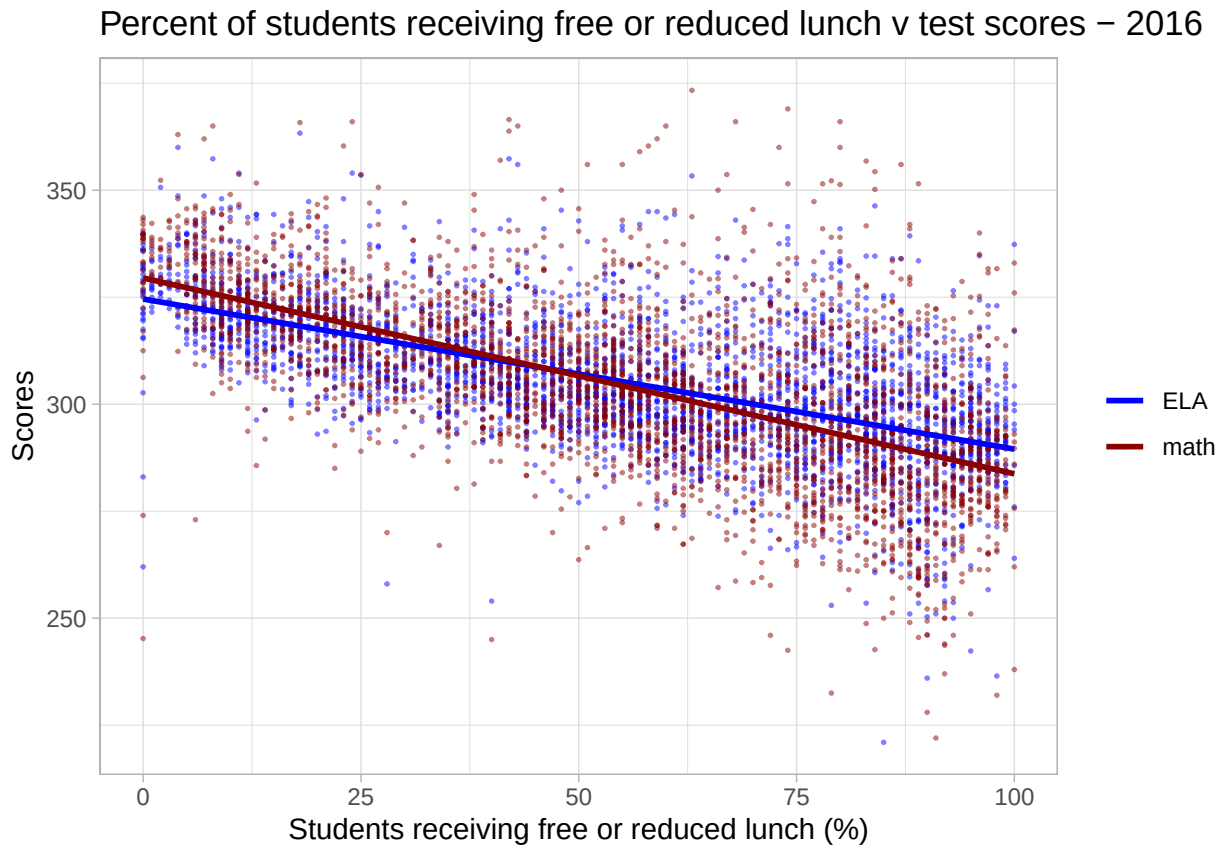
```
# set all ggplots to the same theme
theme_set(theme_light())
```

Create a general scatter plot function to re-use later

```
# create a function called `scatter_plot` that takes data and a mapping as parameters
# the function only plots observations in the data with no missing data
# this should not be a problem since the future code only selects relevant columns
# the plot contains points, trend lines, and a simplified legend
scatter_plot <- function(plot_data, map){
  ggplot(plot_data[complete.cases(plot_data), ], mapping = map) +
    geom_point(size = 0.3, alpha = 0.5) +
    geom_smooth(method = "lm", se = FALSE) +
    theme(legend.title = element_blank())
}
```

Plot the relationship between access to free/reduced price lunch and test performance for 2016. Each point corresponds to a school.

```
# `model_data` contains a subset of the original schools data
# convert percentage to basis points for easy interpretation
model_data <- schools[year == max(year), .(school_name, school_cd, mean_ela_score, mean_math_score, per

# reshape the data from wide to long for plotting purposes
plot_data <- melt(model_data, id.vars = c("school_name", "school_cd", "per_free_reduced_lunch"))

scatter_plot(plot_data, aes(x = per_free_reduced_lunch, y = value, color = variable)) +
  scale_color_manual(values = c("blue", "darkred"), labels = c("ELA", "math")) +
  labs(title = "Percent of students receiving free or reduced lunch v test scores - 2016",
       x = "Students receiving free or reduced lunch (%)",
       y = "Scores")
```



Is free/reduced lunch a significant predictor of ELA scores?

```
model = lm(formula = mean_ela_score ~ per_free_reduced_lunch, data = model_data)
summary(model)
```

```
##
## Call:
## lm(formula = mean_ela_score ~ per_free_reduced_lunch, data = model_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.765  -7.688  -1.004   6.780  51.218
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            324.57244    0.48604  667.79   <2e-16 ***
## per_free_reduced_lunch  -0.35068    0.00784  -44.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.69 on 3350 degrees of freedom
##   (214 observations deleted due to missingness)
## Multiple R-squared:  0.3739, Adjusted R-squared:  0.3738
## F-statistic:  2001 on 1 and 3350 DF,  p-value: < 2.2e-16
```

Is free/reduced lunch a significant predictor of math scores?

```r
model = lm(formula = mean_math_score ~ per_free_reduced_lunch, data = model_data)
summary(model)
```

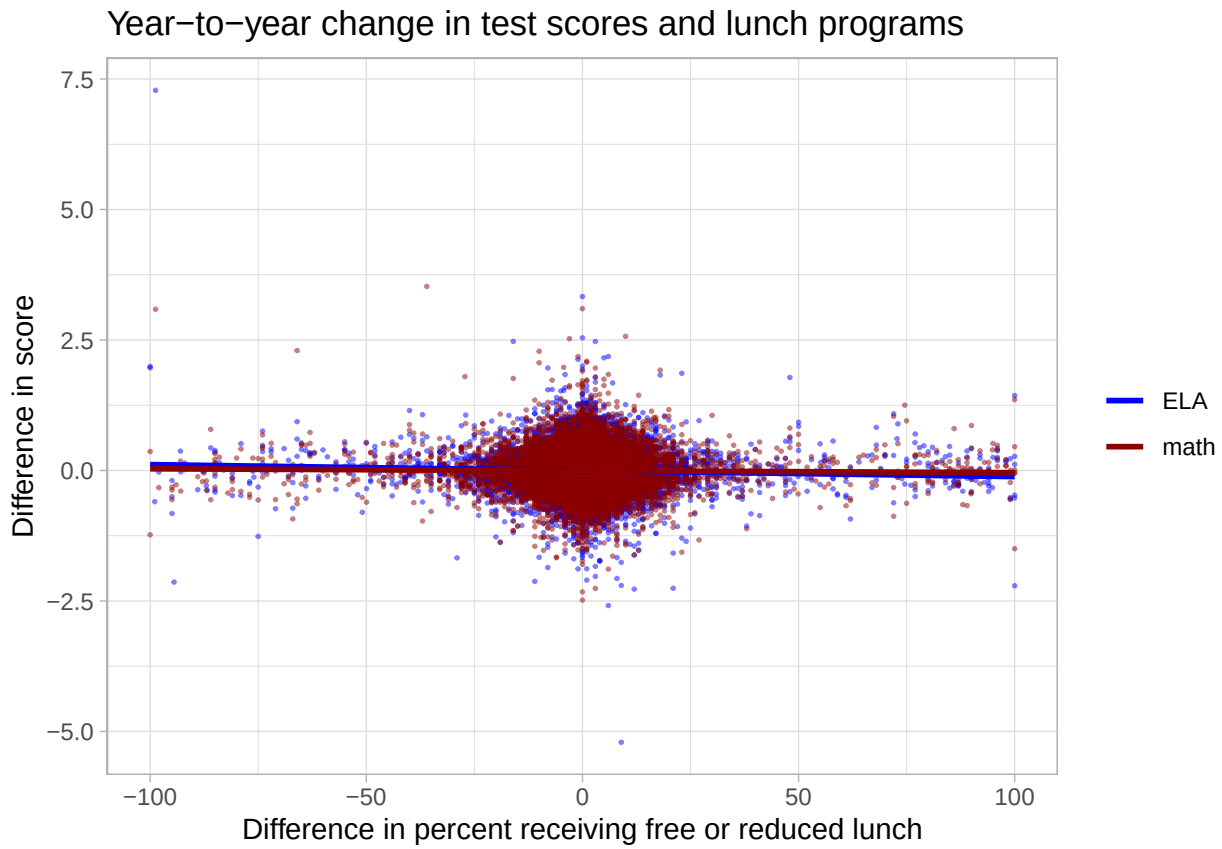
```
##
## Call:
## lm(formula = mean_math_score ~ per_free_reduced_lunch, data = model_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.245  -9.462  -0.895   8.138  73.366
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             329.495466   0.611755  538.61   <2e-16 ***
## per_free_reduced_lunch   -0.457586   0.009868  -46.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.97 on 3351 degrees of freedom
##   (213 observations deleted due to missingness)
## Multiple R-squared:  0.3909, Adjusted R-squared:  0.3907
## F-statistic:  2150 on 1 and 3351 DF,  p-value: < 2.2e-16
```

> In 2016, for every basis point increase in students enrolled in free or reduced lunch, ELA scores decrease by 0.35 points and math scores decrease by 0.45 points.

This does not mean that free and reduced lunch programs **cause** a decrease in scores; rather students who need the free and reduced lunch programs tend to have lower scores. Does a *change* in the use of programs result in a change of scores?

Let's compare the difference in use of lunch programs with difference in scores from year to year.

```r
# create a copy of `schools` so that we preserve the original data
# convert percentage to basis points for easy interpretation
model_data <- schools[, .(year, school_name, school_cd, z_mean_ela_score, z_mean_math_score, per_free_r

# order `schools_plot` by school and year before creating difference variables
# because we are comparing scores year by year, we have to use the z-scores instead of the actual score
setorder(model_data, school_cd, year)
model_data[, diff_lunch := per_free_reduced_lunch - shift(per_free_reduced_lunch), by = school_cd]
model_data[, diff_ELA_score := z_mean_ela_score - shift(z_mean_ela_score), by = school_cd]
model_data[, diff_math_score := z_mean_math_score - shift(z_mean_math_score), by = school_cd]

# reshape the data from wide to long for plotting multiple categories
plot_data <- melt(model_data[, .(school_name, school_cd, year, diff_lunch, diff_ELA_score, diff_math_sc
                  id.vars = c("school_name", "school_cd", "year", "diff_lunch"))

scatter_plot(plot_data, aes(x = diff_lunch, y = value, color = variable)) +
  scale_color_manual(values = c("blue", "darkred"), labels = c("ELA", "math")) +
  labs(title = "Year-to-year change in test scores and lunch programs",
       y = "Difference in score",
       x = "Difference in percent receiving free or reduced lunch")
```



Is a change in free/reduced lunch a significant predictor a change in ELA scores?

```
model = lm(formula = diff_ELA_score ~ diff_lunch, data = model_data)
summary(model)
```

```
##
## Call:
## lm(formula = diff_ELA_score ~ diff_lunch, data = model_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1951 -0.1897 -0.0048  0.1849  7.1721
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0020818  0.0019013  -1.095    0.274
## diff_lunch  -0.0011447  0.0002017  -5.675  1.4e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3267 on 29680 degrees of freedom
##   (5962 observations deleted due to missingness)
## Multiple R-squared:  0.001084,   Adjusted R-squared:  0.00105
## F-statistic: 32.21 on 1 and 29680 DF,  p-value: 1.397e-08
```

Is a change in free/reduced lunch a significant predictor a change in math scores?

```r
model = lm(formula = diff_math_score ~ diff_lunch, data = model_data)
summary(model)
```

```
##
## Call:
## lm(formula = diff_math_score ~ diff_lunch, data = model_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4791 -0.1845 -0.0032  0.1778  3.5176
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0045332  0.0018215  -2.489   0.0128 *
## diff_lunch  -0.0003728  0.0001932  -1.930   0.0537 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.313 on 29674 degrees of freedom
##   (5968 observations deleted due to missingness)
## Multiple R-squared:  0.0001255,  Adjusted R-squared:  9.176e-05
## F-statistic: 3.723 on 1 and 29674 DF,  p-value: 0.05367
```

> For every basis point increase in students enrolled in free or reduced lunch from year-to-year, ELA scores decrease by 0.11 points and math scores decrease by 0.04 points. The effect on ELA scores is significant, but the effect on math scores is not.

There are two possible drivers for the change in program enrollment

1. An increased **access** of the program to students in need. If scenario is true, then our results show that an increase in access does not help to increase test scores
2. An increased number of students in need. If scenario is true, then our results only show us what we already know from the previous exercise: schools with a large percentage of students in free or reduced lunch programs tend to have lower test scores. Unfortunately this does not tell us whether the accessibility of the lunch program has an effect on test scores.

Unfortunately, without more granular data, we do not know which scenario is more accurate (could be both!).

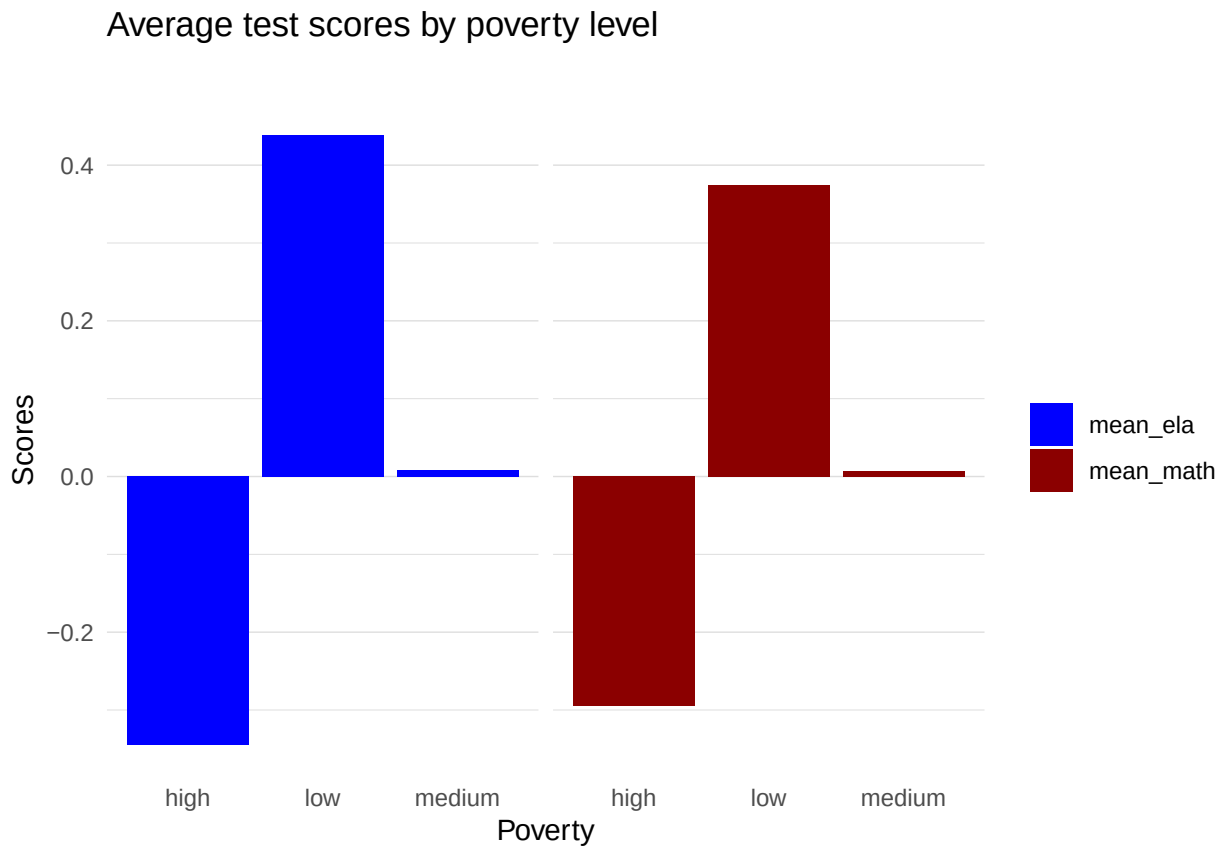
To diagnose which scenario is more accurate, we would want to know whether the same students are represented in the data from year to year and whether their ability to afford lunch stayed constant from year to year.

We assume an increased access to affordable lunch has an effect on test scores over the same year, however, it may be true that an effect on test scores does not show up until *more* than a year of continued access to affordable lunch for students in need.

4.2 Data visualization and analysis (poverty rate)

Average test performance across *counties* with high, low, and medium poverty.

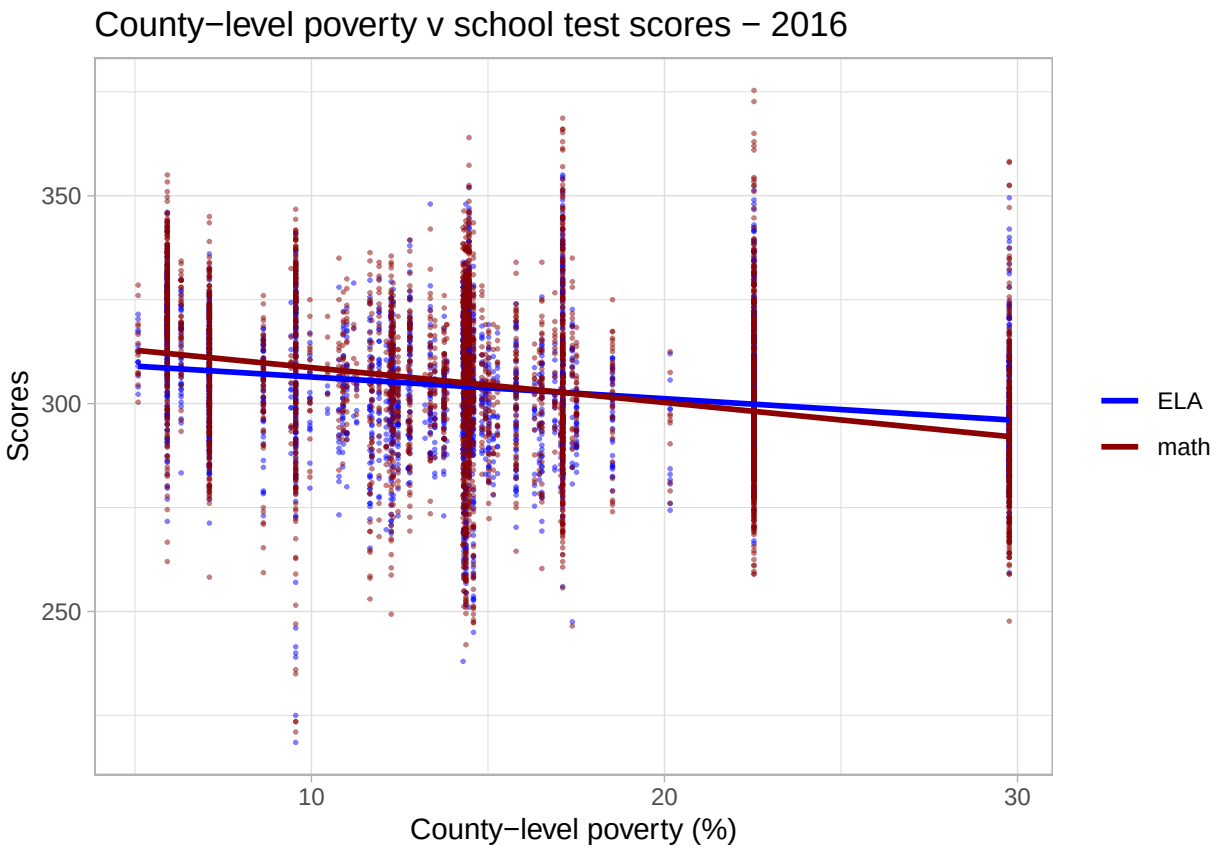
```
# `plot_data` contains a subset of the original merged data
plot_data <- merged[ , .(mean_ela = mean(z_mean_ela_score, na.rm = TRUE),
                         mean_math = mean(z_mean_math_score, na.rm = TRUE)), pov_cat]

# reshape the data from wide to long for plotting purposes
plot_data <- melt(plot_data, id.vars = "pov_cat")
# remove missing data
plot_data <- plot_data[complete.cases(plot_data), ]

ggplot(plot_data[!is.na(pov_cat)]) +
  geom_col(aes(x = pov_cat, y = value, fill = variable)) +
  scale_fill_manual(values = c("blue", "darkred")) +
  facet_grid( ~variable) +
  labs(title = "Average test scores by poverty level", y = "Scores", x = "Poverty") +
  theme(strip.background = element_blank(),
        panel.border = element_blank(),
        axis.ticks = element_blank(),
        panel.grid.major.x=element_blank(),
        legend.title = element_blank())
```



There is a clear difference between the scores from high poverty counties compared to the scores from low poverty counties. Let's again use a scatter plot to visualize the relationship.

```
# `model_data` contains a subset of the original merged data
# convert percentage to basis points for easy interpretation
model_data <- merged[year == 2016, .(school_name, school_cd, mean_ela_score, mean_math_score, county_pe
```

```r
# reshape the data from wide to long for plotting purposes
plot_data <- melt(model_data, id.vars = c("school_name", "school_cd", "county_per_poverty"))

scatter_plot(plot_data, aes(x = county_per_poverty, y = value, color = variable)) +
  scale_color_manual(values = c("blue", "darkred"), labels = c("ELA", "math")) +
  labs(title = "County-level poverty v school test scores - 2016",
       x = "County-level poverty (%)",
       y = "Scores")
```



The two groups on the far right with the highest levels of poverty are Bronx County and Kings County (Brooklyn).

```r
model = lm(formula = mean_ela_score ~ county_per_poverty, data = model_data)
summary(model)
```

```
##
## Call:
## lm(formula = mean_ela_score ~ county_per_poverty, data = model_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -88.14  -9.92  -0.17  10.22  53.42
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)        311.63452    0.69964   445.4   <2e-16 ***
## county_per_poverty  -0.52252    0.04212   -12.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.91 on 3358 degrees of freedom
##   (240 observations deleted due to missingness)
## Multiple R-squared:  0.04382,    Adjusted R-squared:  0.04353
## F-statistic: 153.9 on 1 and 3358 DF,  p-value: < 2.2e-16
```

There is a significant relationship between poverty rates and test scores. However, we see a very poor fit (low R2). It looks like there is a large variance of average scores across schools in each county. Confirm whether this is true in the data.

```
# create a summary table with the standard deviations of scores across counties
# include the mean of scores and the total enrollment across counties for reference
score_sd <- merged[year == 2016, .(mean_ela = mean(mean_ela_score, na.rm = TRUE),
                                   mean_math = mean(mean_math_score, na.rm = TRUE),
                                   std_ela = sd(mean_ela_score, na.rm = TRUE),
                                   std_math = sd(mean_math_score, na.rm = TRUE),
                                   total_enroll = sum(total_enroll, na.rm = TRUE),
                                   county_per_poverty = mean(county_per_poverty)), county_name]

# low poverty schools:
score_sd[order(county_per_poverty)] %>% slice(1:10)
```

```
##     county_name mean_ela mean_math   std_ela  std_math total_enroll
##  1:      PUTNAM 312.0060  313.3393  5.766324  7.979840         9433
##  2:      NASSAU 316.8058  320.1810 13.399378 16.768997       133773
##  3:    SARATOGA 312.2846  318.3476 10.243646 10.235388        21850
##  4:     SUFFOLK 306.5373  307.6212 12.001177 15.210646       156732
##  5:    DUTCHESS 302.8058  301.7847 11.310833 15.156105        24365
##  6:       ESSEX 304.6819  306.3194  9.897800 11.887405         2909
##  7: WESTCHESTER 307.4006  308.1686 21.244743 25.613503       105793
##  8:     ONTARIO 301.1042  304.7031 10.324705 12.449564        11628
##  9:     WYOMING 302.3472  310.5694  3.923723  6.793055         2085
## 10:      CAYUGA 296.2611  306.7667 13.752448 13.029469         5727
##     county_per_poverty
##  1:         0.05091140
##  2:         0.05917006
##  3:         0.06309991
##  4:         0.07108796
##  5:         0.08638583
##  6:         0.09425359
##  7:         0.09553986
##  8:         0.09957058
##  9:         0.10456122
## 10:         0.10781514
```

```
# high poverty schools:
score_sd[order(county_per_poverty, decreasing = TRUE)] %>% slice(1:10)
```

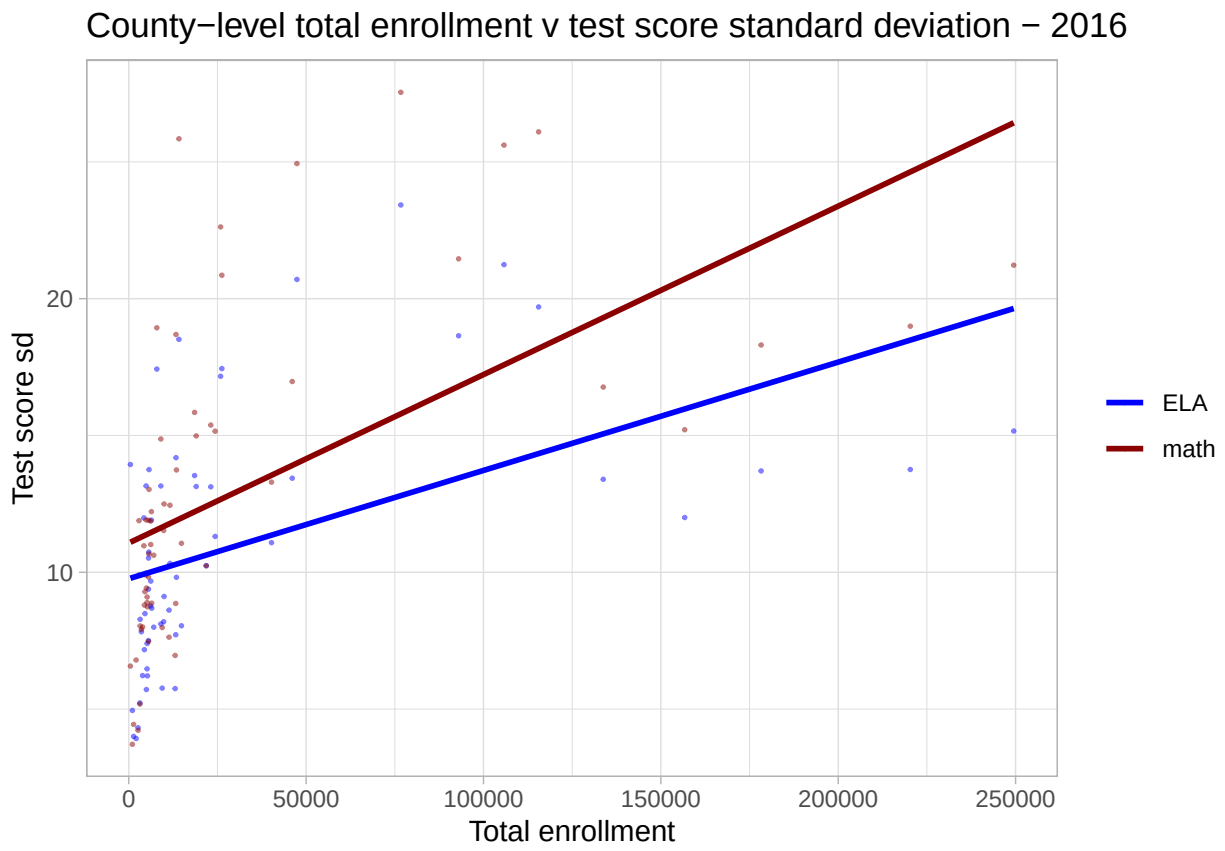
```
##        county_name mean_ela mean_math   std_ela  std_math total_enroll
```

```
##  1:          BRONX 295.7927  291.0076 13.704320 18.308290       178274
##  2:          KINGS 303.6666  300.7693 15.162816 21.224997       249479
##  3:     MONTGOMERY 289.5083  293.2167 13.156375 11.910948         4903
##  4:     CHAUTAUQUA 297.4485  300.4387  9.816027 13.740168        13445
##  5:         OSWEGO 294.3958  302.3482  5.748932  6.959837        13082
##  6:       TOMPKINS 302.0362  306.2826 17.427408 18.936476         7916
##  7: SAINT LAWRENCE 298.2115  301.6731  8.617335  7.627652        11368
##  8:       NEW YORK 307.3281  306.6173 19.698068 26.097873       115541
##  9:       FRANKLIN 290.0714  291.6964  6.470170  8.895981         5177
## 10:    CATTARAUGUS 300.2958  306.5883  8.108884 10.125134         9051
##     county_per_poverty
##  1:          0.2976889
##  2:          0.2253745
##  3:          0.2016228
##  4:          0.1852780
##  5:          0.1750848
##  6:          0.1739076
##  7:          0.1713187
##  8:          0.1711614
##  9:          0.1706825
## 10:          0.1689680
```

In particular, the Bronx, which has the highest poverty rate, there is an average of 13 point variation from the mean across schools' ELA test scores, and an average of 18 point variation from the mean across schools' math test scores.

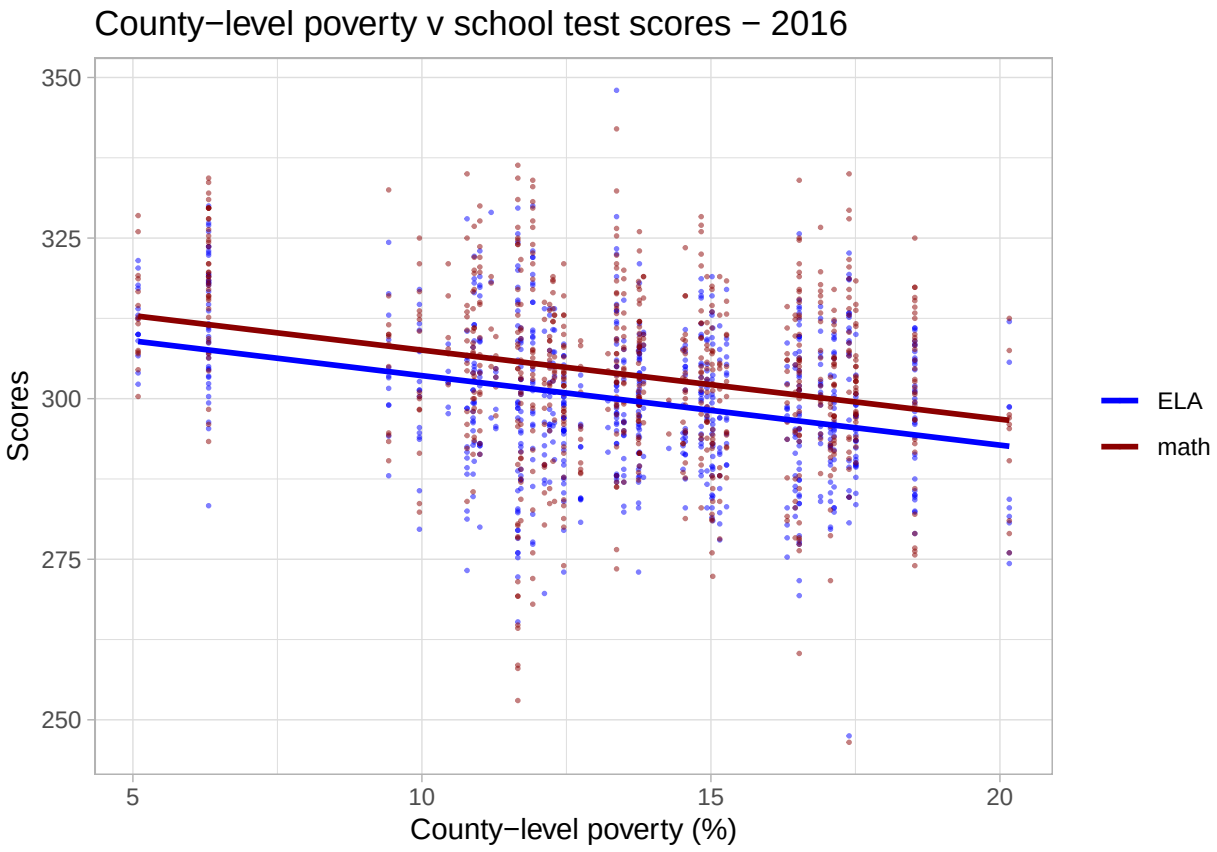
Across all schools, math scores have a higher variance than ela scores. There doesn't seem to be a difference in variance across poverty levels. Rather, variance has more to do with total enrollment. This makes sense - counties with a large number of students are likely to have more diversity in test results. Notably Kings County (Brooklyn) and New York County (Manhattan) both have very large variances.

```
# reshape the data from wide to long for plotting purposes
plot_data <- melt(score_sd[,.(county_name, std_ela, std_math, total_enroll)], id.vars = c("county_name"

scatter_plot(plot_data, aes(x = total_enroll, y = value, color = variable)) +
  scale_color_manual(values = c("blue", "darkred"), labels = c("ELA", "math")) +
  labs(title = "County-level total enrollment v test score standard deviation - 2016",
       x = "Total enrollment",
       y = "Test score sd")
```




```
low_enroll_counties <- score_sd[total_enroll <= quantile(score_sd$total_enroll, 0.75), .(county_name)]
```

```
# `model_data` contains a subset of the original merged data
# convert percentage to basis points for easy interpretation
model_data <- merged[year == 2016, .(county_name, school_name, school_cd, mean_ela_score, mean_math_sco
# keep only counties with enrollment below the 75th percentile
# after the merge, we have no use for `county_name`
model_data <- model_data[low_enroll_counties, on = "county_name"][, county_name := NULL]


# reshape the data from wide to long for plotting purposes
plot_data <- melt(model_data, id.vars = c("school_name", "school_cd", "county_per_poverty"))
scatter_plot(plot_data, aes(x = county_per_poverty, y = value, color = variable)) +
  scale_color_manual(values = c("blue", "darkred"), labels = c("ELA", "math")) +
  labs(title = "County-level poverty v school test scores - 2016",
       x = "County-level poverty (%)",
       y = "Scores")
```




```
model = lm(formula = mean_ela_score ~ county_per_poverty, data = model_data)
summary(model)
```

```
##
## Call:
## lm(formula = mean_ela_score ~ county_per_poverty, data = model_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.087  -7.181  -0.444   7.487  48.059
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        314.4063     1.7279 181.956   <2e-16 ***
## county_per_poverty  -1.0821     0.1237  -8.745   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.1 on 780 degrees of freedom
##   (48 observations deleted due to missingness)
## Multiple R-squared:  0.08929,    Adjusted R-squared:  0.08812
## F-statistic: 76.47 on 1 and 780 DF,  p-value: < 2.2e-16
```

By removing schools from counties with large enrollment, our R2 has increased slightly from 0.04353 to 0.08812. This still means that **poverty rate only captures about 8% of variation in average test**

scores. Because of this large variation in test scores across schools, it makes more sense to use a school-level variable like percentage of students with free or reduced lunch as a proxy for measuring poverty, as opposed to aggregating data for all schools across counties.

Bonus: does percentage of students with free and reduced lunch capture the variation in county poverty?

```
# subset from `merged`, aggregate to country level
sum_table <- merged[ , .(tot_enroll = sum(total_enroll, na.rm = TRUE),
                         tot_reduced_lunch = sum(num_reduced_lunch, na.rm = TRUE),
                         tot_free_lunch = sum(num_free_lunch, na.rm = TRUE),
                         county_per_poverty = mean(county_per_poverty, na.rm = TRUE))
                     , county_name]

# convert totals to percentages
sum_table[, per_free_reduced_lunch := (tot_reduced_lunch + tot_free_lunch) * 100/ tot_enroll]

# remove the columns with total student numbers
sum_table[, tot_enroll := NULL][, tot_reduced_lunch := NULL][, tot_free_lunch := NULL]
head(sum_table)
```

```
##    county_name county_per_poverty per_free_reduced_lunch
## 1:      ALBANY          0.1229838               38.78464
## 2:    ALLEGANY          0.1508549               51.80581
## 3:       BRONX          0.2872294               85.85094
## 4:      BROOME          0.1584756               47.49617
## 5: CATTARAUGUS          0.1642522               50.31271
## 6:      CAYUGA          0.1148119               42.47888
```

```
model = lm(formula = county_per_poverty ~ per_free_reduced_lunch, data = sum_table)
summary(model)
```

```
##
## Call:
## lm(formula = county_per_poverty ~ per_free_reduced_lunch, data = sum_table)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.061575 -0.012958  0.000547  0.011040  0.062058
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             0.0083820  0.0105742   0.793    0.431
## per_free_reduced_lunch 0.0026707  0.0002239  11.930   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02051 on 60 degrees of freedom
## Multiple R-squared:  0.7034, Adjusted R-squared:  0.6985
## F-statistic: 142.3 on 1 and 60 DF,  p-value: < 2.2e-16
```

> 70% of the variation in county poverty level is captured by the average enrollment in free and reduced lunch programs across schools.