Research on how New York City schools' demographic influence the student's chronical absent rate

Liangjiayi Wang

ECO225 2020.09-2020.12

Student# 1004405789

Dec 18. 2020

Introduction:

Nowadays, more and more students in schools are increasingly impacted by chronic absent rate. Chronic absenteeism is defined as student miss at least 15 days of school in a year—are at serious risk of falling behind in school. Student chronic absent rate is an important an important measure of educational success. (Brendan Bartanen, 2020) For instance, we can use it to predict the graduation rates, school readiness and enrollment number. According to the 2015-16 Civil Rights Data Collection, over 7 million students, corresponding to 16 percent of the student population—or about 1 in 6 students, missed 15 or more school days in 2015-16. (U.S. Department of Education, 2019). In other words, there were about more than 100 million school days lost, and this excessive amount of absenteeism phenomenon was a nationwide crisis.

While in New York City, the rate of chronic absenteeism, around 26 percent, is much higher compare to U.S. average rate. (Tina Posterli,2018) It is worth reviewing the causes for chronic absenteeism. Based on this, we can help all students have a better chance of reaching their full potential, and the government and the schools can know how to reduce the students' long-term chronic absent rate. In existing literature review, researchers mainly focus on how chronic absence influence the Student Achievement. Students who are repeatedly absent from school, caused the missing of important learning and developmental opportunities, are more likely to have negative consequences on their future outcomes such as difficulties in academic achievement, learning, sociability, and mental health. (London, Sanchez, & Castrechini, 2016) However, we cannot find any research that builds a statistical model to measure the level of influence of each independent variable on the absent rate.

Our research will mainly research how New York City schools' demographics influence students'

chronically absent rate. We can use those variables to answer three key questions: (1) What percentage of students are chronically absent in 2016? (2) How each school environment and characteristics variables influence the chronically absent rate? (3) How much of those independent variables impact the chronic absent rate?

In the following analysis, we will use the summary table, boxplot, bar graph, and correlation table to illustrate better the factors that affect the student's absence and visualize our initial findings. Then, we will build the statistical model based on the variables we have in previous part and see at what level those independent variables impact the chronic absent rate. In this part, firstly, we will build a model to find how the percent of chronically absent rate varies with one specific variable (the Economic Need Index) while holding other factors stays the same. Considering Economic Need Index as one main effect, we will add other independent variables like Races, Population, and Collaborative Teachers % into the model. In general, the ideal model is adding two characteristics variables (e.g., Race, Population), and two environment variables (e.g., Collaborative Teachers %, Economic Need Index). And we may adjust our model based on the model selection results.

We will use the R-squared statistic to compare regression models. A larger R-squared value model means that a larger percentage of the variation in the dependent variable that is explained by independent variables. Thus, we expect that the R-square value increases as more variables add to the model, which better predicts the chronically absent rate. If the R-squared value is the same for the two models, we will consider finding the model with the lowest value of the Akaike Information Criterion (AIC) and Schwarz' Bayesian Information Criterion (BIC). AIC and BIC are both penalized-likelihood criteria.

Data Sources:

In this research, our data is gathered from Kaggle of Data Science for Good: PASSNYC (https://www.kaggle.com/passnyc/data-science-for-good). PASSNYC is a not-for-profit organization dedicated to promoting educational opportunities for New York City's talented and underserved students. In 2016, PASSNYC collected 1273 schools' data to identify students within New York City's under-performing school districts and aims to increase the diversity of students taking the Specialized High School Admissions Test (SHSAT). The 2016 School Explorer dataset contains 1273 New York schools and each school's specific demographics like absent rate, races distribution, location and collaborative teachers rating etc.

Using PASSNYC (2016 School Explorer) dataset, we can measure schools' performance in education. In this dataset, we mainly use the location variables like school name, city, longitude, latitude, and zip; the characteristic variables like Percent Asian, Percent Black, Percent Hispanic, and Percent White (note: those variables represent the percentage of each race in each school and independently scale from 0%-100%); the environment variables like Economic Need Index and Collaborative Teachers %; And the unique dependent variables: Percent of Students Chronically Absent. We convert all the variables with percentage to a fraction (i.e. 8% into 0.08). Meanwhile, there are more than 40 areas of New York City; we will select 7 areas(Brooklyn, Bronx, New York, Staten Island, Jamaica, Flushing, Long Island City) with the greatest number of schools of New York City and then look at the demographic variables of each schools within 7 areas to analyze their influence on absent rate. After remove missing values, the dataset now contains 1059 rows × 12 columns.

Moreover, the population is one of the key figures for demographics which can measure an area's

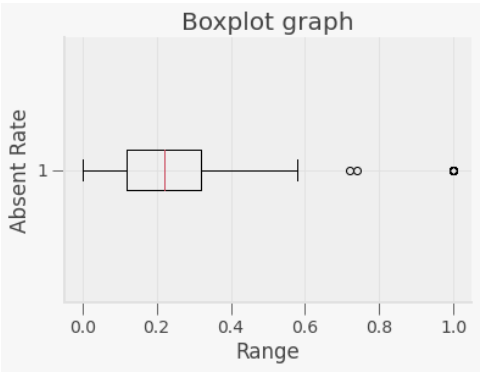
prosperity and infrastructure level, and have impact on students' behavior. Thus, we also want to find the relationship between the truancy rate and population. The population data can be gathered from website: New York Zip Codes by Population (https://www.newyork-demographics.com/zip_codes_by_population) which record by the US Census. It included 1753 New York zip codes and collected accurate population information based on the zip code in New York City. Based on this data, I can research if there exists a linear relationship between the population and the absence rate. Since our original data and the new data contain columns with zip code information, we can easily merge two datasets by zip and add the population information into the original data.

Data Visualization and Summary Statistics:

First, we compute the descriptive statistics to summarize the central tendency, dispersion and shape of Absent Rate, Economic Need Index, and Collaborative Teachers % distribution. And we also draw a boxplot graph which provides a graphical summary of the distribution of a sample.

Table 1: Summary statistics of chronic absent rate, Economic Need Index, and Collaborative Teachers %

	Mean	Std	Min	25%	50%	75%	Max
Chronic Absent Rate	0.23	0.14	0.00	0.12	0.22	0.32	1.00
Economic Need Index	0.70	0.20	0.05	0.60	0.76	0.85	0.96
Collaborative Teachers %	0.88	0.08	0.00	0.84	0.90	0.94	1.00



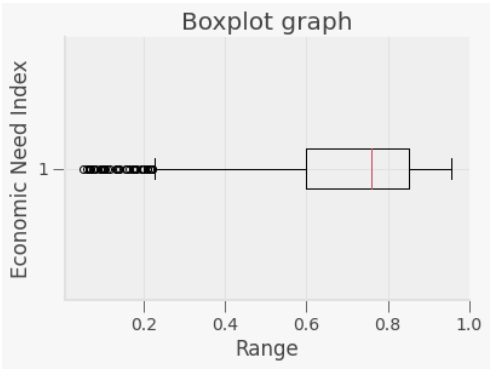


Figure 1(left): Boxplot graph for absent rate *Figure 2(right): Boxplot graph for Economic Need Index*

We can use that information from table 1 to answer the first question: What percentage of students are chronically absent in 2016? For all schools within 7 areas, the mean (the average of the data) of the absent rate is 0.23. The mean for Economic Need Index is 0.70, and the Collaborative Teachers % is 0.88. The Absent Rate/ Collaborative Teachers % scale from 0 to 1, and the Economic Need Index scale from 0.049 to 0.957. We have a large mean and median in variables Collaborative Teachers %. This observation a good indicator because it shows that most schools have a great percentage of collaborative teachers, which help student engagement.

From the figure 1, we find that the data is right-skewed distribution and it is more concentrate on 0 to 0.4. From the figure 2, we find that the distribution of Economic Need Index is not uniform, and there exists several extremely small values(outliers).

We predict that the area with more schools may have larger absent rate since the sample size for those areas is larger (i.e. More students); meanwhile, the ethnic backgrounds of students are more diverse in those areas. We use the aggregate and group-by function to see 7 areas schools' performance on the absent rate.

City	amin	amax	mean	median	std
BRONX	0.00	1.00	0.278144	0.280	0.121737
BROOKLYN	0.00	1.00	0.221546	0.200	0.139872
FLUSHING	0.02	0.29	0.099000	0.090	0.062442
JAMAICA	0.01	0.40	0.214194	0.220	0.098920
LONG ISLAND CITY	0.00	0.45	0.152381	0.130	0.102757
NEW YORK	0.00	1.00	0.217111	0.180	0.181362
STATEN ISLAND	0.05	0.41	0.182333	0.155	0.084620

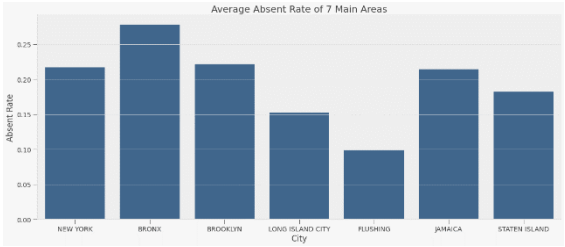


Figure 4(Left): Summary statistics of Absent Rate of 7 Main Areas

Figure 5(Right): Bar graph - Average Absent Rate of 7 Main Areas

From the summary statistics table, we noticed that Flushing has the smallest aggregate max absent rate and smallest mean. The Bronx has the largest mean and median. Long Island City and Staten Island's performance are quite similar, and they have a small rate range. The bar graph visualizes the absent student rate is highest in Bronx (with mean 0.28), led by Brooklyn (with mean 0.22) and New York (with mean 0.21). Jamaica (with mean 0.20), Staten Island (with mean 0.18). Flushing (with mean 0.01) and Long Island City (with mean 0.13) have a relatively low absent rate.

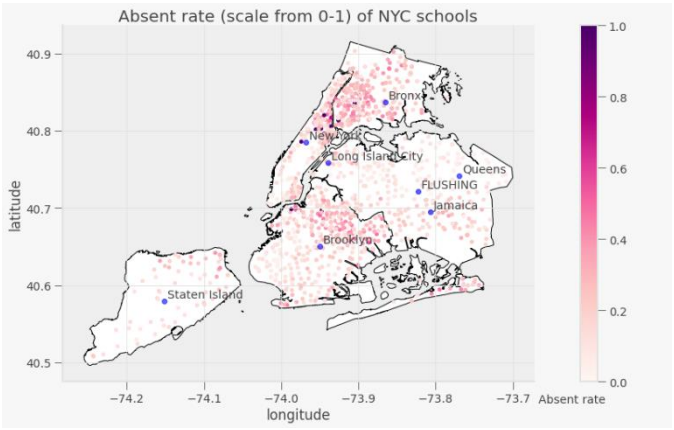


Figure 6: Map - Absent rate (scale from 0-1) of NYC schools

In the map, the pink points reflect the absent rate of schools in NYC. The darker points appear, the greater the absent rate. I label the 7 main areas' location by blue points. As we discussed in the last part, New York, Bronx and Brooklyn have a serious problem with students' absent rates. Moreover, those 3 areas have the most significant number of schools. It matches our assumption that the area with more schools will have the larger absent rate.

To find the second question: How each school environment and characteristics variables influence the chronically absent rate? We create a correlation matrix to find the correlation between variables.

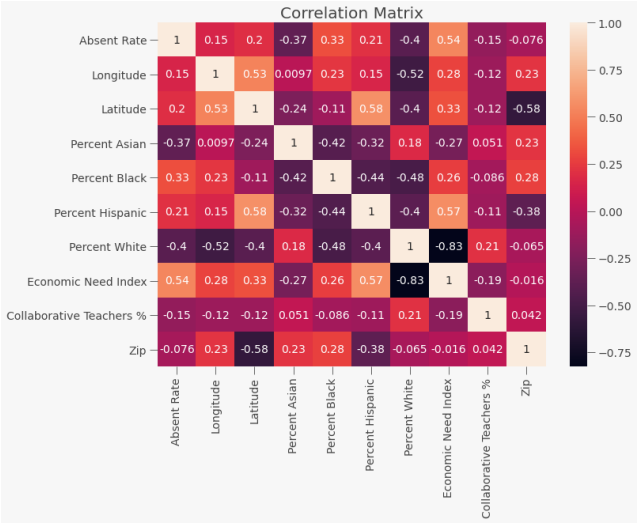


Figure 6: Heat map - Correlation Matrix

Absent rate is negatively correlated with Percent Asian (-0.37) and Percent White (-0.4). However, Absent rate is positively correlated with Percent Black (0.33) and Percent Hispanic (0.21). In other words, schools with large percent of Black and Hispanic have greater absent rates. In other words, schools with large percent of Black and Hispanic may have greater absent rates.

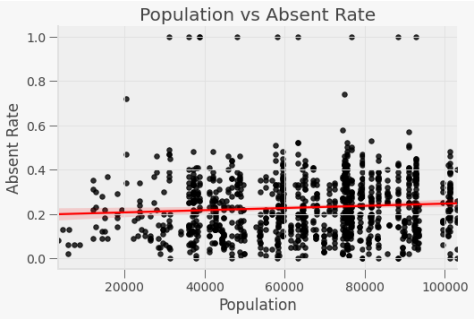
The Collaborative Teachers % are negatively correlated with absent rate.

We use 2016 School Explorer to see how the school's location, race percentage, Collaborative Teachers % and economic need index affect the absent rate based on the summary statistics, visualized graphs and map. After the web scraping process, we add one more characteristic variables-Population into our dataset. We will build a new relationship between population and absent rate. The hypothesis is that if an area with a large population, that area's absence rate will be much greater. We will use the summary statistics table to have an overall idea about the population variable.

Table 2: Summary statistics of Population

	Mean	Std	Min	25%	50%	75%	Max
Population	65222.78	22168.01	3335.00	46843.00	67948.00	81716.00	103123.00

The population has the close mean (65222.78) and median (67948.00). The distribution of data looks uniform. Then we create a regression plot.



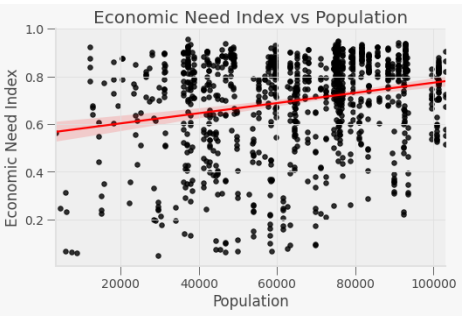


Figure 7(left): Regression Plot-Population vs Absent Rate

Figure 8(right): Regression Plot-Economic Need Index vs Population

In figure 7, we find that the direct effects of population on absence rate are inapparent. Thus, I do not support the hypothesis that if an area with a large population, that area's absence rate will be much greater. In figure 8, we have a much steeper slope, which proves that the population's influence on the Economic Need Index is apparent. There may exist an endogeneity problem with the Population and Economic Need Index. The population may indirectly affect the absent rate, which acts based on the population's effects on the Economic Need Index. In other words, the increase in population leads to the rise in Economic Need Index, and then the increase in Economic Need Index causes an increase in the absent rate. We will do a future study of the relationship between Population vs. Absent Rate in the modelling part.

Regression Results:

To check whether our Y (Absent rate) and X (Economic Need Index)'s economic relationship is linear, we start by creating a scatterplot. We study the overall pattern of the plotted points.

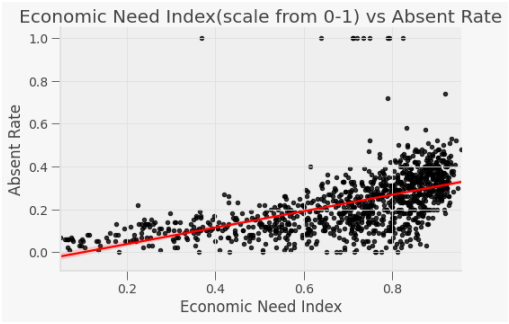


Figure 9: Regression Plot-Economic Need Index vs Absent Rate

The plot shows a reasonably positive linear relationship between Economic Need Index and Absent Rate since there exists an upward slope and a straight-line pattern in the plotted data points. Also, in last part, we noticed that the correlation between those two variables are 0.54. Specifically, if higher Economic Need Index students in school, the chronically absent rate will be much higher. Given the plot, choosing a linear model to describe this relationship seems like a reasonable assumption.

Ordinary least squares (OLS) regression helps to estimate the relationship between one or more independent variables and a dependent variable. In this case, the independent variable is Economic Need Index and dependent variable is Chronic Absent Rate.

OLS Regression Results

Dep. Variable:	Chronic_Absent_Rate	R-squared:	0.293
Model:	OLS	Adj. R-squared:	0.292
Method:	Least Squares	F-statistic:	437.6
Date:	Sat, 19 Dec 2020	Prob (F-statistic):	1.36e-81
Time:	05:45:14	Log-Likelihood:	730.75
No. Observations:	1059	AIC:	-1457.
Df Residuals:	1057	BIC:	-1448.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0390	0.013	-2.926	0.004	-0.065	-0.013
Economic_Need_Index	0.3831	0.018	20.920	0.000	0.347	0.419

Omnibus:	612.953	Durbin-Watson:	1.431
Prob(Omnibus):	0.000	Jarque-Bera (JB):	9647.819
Skew:	2.331	Prob(JB):	0.00
Kurtosis:	17.032	Cond. No.	7.37

Figure 10: OLS Regression Result of Model1 -Economic Need Index vs Absent Rate

We can write our simple linear regression model as

$$ChronicAbsentRate_i = \beta_0 + \beta_1 * EconomicNeedIndex_i + u_i$$

where:

- β_0 is the intercept of the linear trend line on the y-axis
- β_1 is the slope of the linear trend line, representing the marginal effect of Economic Need Index against Chronic Absent Rate
- u_i is a random error term (deviations of observations from the linear trend due to factors not included in the model)

From our results, we see that

- The intercept $\hat{\beta}_0 = -0.0390$
- The slope $\hat{\beta}_1 = 0.3831$
- The positive $\hat{\beta}_1$ parameter estimate implies that Economic Need Index has a positive effect on economic outcomes, as we saw in the figure.
- The p-value of 0.000 for $\hat{\beta}_1$ implies that the effect of Economic Need Index is statistically significant (using $p < 0.05$ as a rejection rule).
- The R-squared value of 0.293 indicates that around 29% of variation in Chronic Absent Rate is explained by Economic Need Index.

This model looks good. And it matches our theory that there exists positive linear relationship between Economic Need Index and Chronic Absent Rate and this effect is noteworthy. In the next step, we will consider the Economic Need Index as one main effect and introduce more explanatory variables to predict a response variable's outcome. More variation is supposed to be explained.

We introduced lots of variables from data visualization part, including Races, Population, and Collaborative Teachers %. And see how they impact the absent rate individually based on the correlation graph and bar graph. Noticed that the economic relationship between those independent variables and chronic absenteeism rate are also linear. Then, we use those variables to answer our final questions: How much of those independent variables impact the chronic absent rate?

Before we fit a multiple linear regression model. In data visualization part, we find a positive linear relationship between population and Economic Need Index. We use the OLS model to check whether there is strong multicollinearity between those two variables.

OLS Regression Results

Dep. Variable:	Chronic_Absent_Rate	R-squared:	0.295
Model:	OLS	Adj. R-squared:	0.294
Method:	Least Squares	F-statistic:	221.0
Date:	Sat, 19 Dec 2020	Prob (F-statistic):	7.08e-81
Time:	00:43:32	Log-Likelihood:	731.51
No. Observations:	1058	AIC:	-1457.
Df Residuals:	1055	BIC:	-1442.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0227	0.016	-1.451	0.147	-0.053	0.008
Economic Need Index	0.3917	0.019	20.816	0.000	0.355	0.429
Population	-3.406e-07	1.73e-07	-1.968	0.049	-6.8e-07	-9.54e-10

Omnibus:	609.038	Durbin-Watson:	1.435
Prob(Omnibus):	0.000	Jarque-Bera (JB):	9484.693
Skew:	2.317	Prob(JB):	0.00
Kurtosis:	16.917	Cond. No.	4.15e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.15e+05. This might indicate that there are strong multicollinearity or other numerical problems.

Figure 11: OLS Regression Result of Model with Endogeneity

The model results indicate that there are strong multicollinearity or other numerical problems. To eliminate the multicollinearity problem, we need to recode the variable Population. Since the Population variables has a range from 3335 to 103123 with median 67948 , we will consider turn it into a categorical variable based on the value of it (value > 67948 : pop_high = 1, value <=67948 : pop_high = 0).We gradually include more variables in the model. The pop_high=1 represents that the area has a significant number of populations.

Then, we gradually include more variables in the model. And see how the regression coefficient

changes and whether the model is improved after we add more variables. We expect R-squared value increase since more independent variables can explain the variation of dependent variables. The coefficient between Economic Need Index and Chronic Absent Rate should remain positive. In other words, we expand the simple regression model into the multiple regression model.

In this way, our model becomes more powerful, and the estimation ability increases. New models can help researchers to earn more information about how different features affect the absence rate. The Bureau of Education can use this data to make policies that reduce the rate of truancy. The decreasing absent race provides more supports for students' future success.

The following are the four models that we may use to predict the absent rate. We will compare them and select the most appropriate model by R-squared, AIC and BIC.

$$Model2: \widehat{ChronicAbsentRate}_i = \beta_0 + \beta_1 EconomicNeedIndex_i + \beta_2 PercentELL_i$$

$$Model3: \widehat{ChronicAbsentRate}_i = \beta_0 + \beta_1 EconomicNeedIndex_i + \beta_2 PercentELL_i + \beta_3 PercentAsian_i + \beta_4 PercentBlack_i + \beta_5 PercentHispanic_i + \beta_6 PercentWhite_i$$

$$Model4: \widehat{ChronicAbsentRate}_i = \beta_0 + \beta_1 EconomicNeedIndex_i + \beta_2 PercentELL_i + \beta_3 PercentAsian_i + \beta_4 PercentBlack_i + \beta_5 PercentHispanic_i + \beta_6 PercentWhite_i + \beta_7 pophigh_i$$

$$Model5: \widehat{ChronicAbsentRate}_i = \beta_0 + \beta_1 EconomicNeedIndex_i + \beta_2 PercentELL_i + \beta_3 PercentAsian_i + \beta_4 PercentBlack_i + \beta_5 PercentHispanic_i + \beta_6 PercentWhite_i + \beta_7 pophigh_i + \beta_8 PercentCollaborativeTeachers_i$$

Table 4-2 - OLS Regressions Summaries for Model 1-5

	Model 2	Model 3	Model 4	Model 5
Intercept	-0.039***	0.497***	0.447**	0.561***
	(0.013)	(0.173)	(0.174)	(0.179)
Percent_Black		-0.605***	-0.552***	-0.560***
		(0.185)	(0.186)	(0.185)
Percent_Hispanic		-0.724***	-0.677***	-0.686***
		(0.184)	(0.184)	(0.184)
Percent_White		-0.486***	-0.430**	-0.429**
		(0.183)	(0.184)	(0.183)
Percent_Collaborative_Teachers				-0.121**
				(0.047)
Economic_Need_Index	0.383***	0.547***	0.562***	0.562***
	(0.018)	(0.037)	(0.037)	(0.037)
Percent_Asian		-0.835***	-0.781***	-0.788***
		(0.183)	(0.184)	(0.184)
pop_high			-0.019**	-0.019***
			(0.007)	(0.007)
R-squared	0.293	0.388	0.391	0.395
R-squared Adj.	0.292	0.385	0.388	0.391
R-squared	0.29	0.39	0.39	0.40
No. observations	1059	1059	1058	1058

Standard errors in parentheses.
* p<.1, ** p<.05, ***p<.01

Figure 12: OLS Regression Result of Model Comparison

There is an increase of *R*-squared as more variables are added. Based on the method we mentioned in the introduction part, we will select model 5 since its *R*-squared value is the largest.

OLS Regression Results

Dep. Variable:	Chronic_Absent_Rate	R-squared:	0.395
Model:	OLS	Adj. R-squared:	0.391
Method:	Least Squares	F-statistic:	97.94
Date:	Sat, 19 Dec 2020	Prob (F-statistic):	4.91e-110
Time:	03:36:23	Log-Likelihood:	812.28
No. Observations:	1058	AIC:	-1609.
Df Residuals:	1050	BIC:	-1569.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.5610	0.179	3.135	0.002	0.210	0.912
Economic_Need_Index	0.5618	0.037	15.209	0.000	0.489	0.634
Percent_Asian	-0.7881	0.184	-4.288	0.000	-1.149	-0.427
Percent_Black	-0.5601	0.185	-3.023	0.003	-0.924	-0.197
Percent_Hispanic	-0.6861	0.184	-3.731	0.000	-1.047	-0.325
Percent_White	-0.4294	0.183	-2.343	0.019	-0.789	-0.070
pop_high	-0.0187	0.007	-2.587	0.010	-0.033	-0.005
Percent_Collaborative_Teachers	-0.1215	0.047	-2.570	0.010	-0.214	-0.029

Omnibus:	686.013	Durbin-Watson:	1.504
Prob(Omnibus):	0.000	Jarque-Bera (JB):	14317.560
Skew:	2.627	Prob(JB):	0.00
Kurtosis:	20.239	Cond. No.	199.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure 13: OLS Regression Result of Selected Model

This model looks good and it avoid endogeneity problem. Thus, our final model can be concluded as

$$\begin{aligned} Model5: \widehat{ChronicAbsentRate_i} = {} & 0.5610 + 0.5618 EconomicNeedIndex_i - 0.7881 PercentAsian_i - 0.5601 PercentBlack_i \\ & - 0.6861 PercentHispanic_i - 0.4294 PercentWhite_i - 0.0187 pophigh_i \\ & - 0.1215 PercentCollaborativeTeachers_i \end{aligned}$$

From our results, we see that $R^2 = 0.395$ indicates that 39.5% of the sample variance is captured in the model. In general, all independent variables are statistically significant since p-value < 0.05.

The percentage of each race, higher population, and Percent Collaborative Teachers will negatively impact Chronic Absent Rate, which is good news. For the characteristic variables, we cannot control students' race and the population around the school, so the smaller the effects of the characteristic variables on the absence rate will help us put more effort into the environment factors that we can make change.

For our two environment variables, we noticed that a higher Percentage of Collaborative Teachers would decrease the absent rate, representing that teachers' behavior will influence students'

engagement. The larger the Economic Need Index leads to a significant increase in the chronic absent rate. Thus, financial support is vital for those schools with high economic index. If a school has sufficient financial support, it will provide a better infrastructure that improves the learning environment and decreases the absence rate.

The following are specific coefficient interpretations:

- Together with Intercept, we can interpret that if the population around school is less than 67948(pop_high=0), holding other variables constant, the Chronic Absent Rate is around 0.5610 on average.
- Interpretation of Economic Need Index coefficient:

Holding other variables constant,1 percentage increase in `Economic Need Index` will lead to $0.5618 * 0.01 \approx 0.006 = 0.6$ percentage increase in absent rate on average.

- Interpretation of Percent Asian coefficient:

Holding other variables constant,1 percentage increase in `Percent Asian` will lead to $0.7881 * 0.01 \approx 0.008 = 0.8$ percentage decrease in absent rate on average.

- Interpretation of Percent Black coefficient:

Holding other variables constant,1 percentage increase in `Percent Black` will lead to $0.5601 * 0.01 \approx 0.006 = 0.6$ percentage decrease in absent rate on average.

- Interpretation of Percent Hispanic coefficient:

Holding other variables constant,1 percentage increase in `Percent Hispanic` will lead to $0.6851 * 0.01 \approx 0.007 = 0.7$ percentage decrease in absent rate on average.

- Interpretation of Percent White coefficient:

Holding other variables constant,1 percentage increase in `Percent White` will lead to $0.4294 * 0.01 \approx 0.004 = 0.4\$$ percentage decrease in absent rate on average.

- Interpretation of pop_high coefficient:

Holding other variables constant, if the population around school is greater than 67948, the `Chronic_Absent_Rate` will decrease by $0.0187 * 0.01 \approx 0.0002 = 0.02$percentage increase in absent rate on average.

- Interpretation of Percent Collaborative Teachers coefficient:

Holding other variables constant,1 percentage increase in `Collaborative Teachers` will lead to $0.1215 * 0.01 \approx 0.002 = 0.2$ percentage decrease in absent rate on average.

Conclusions and next steps:

This literature mainly researches how New York City schools' demographics influence students' chronically absent rate. Unlike other literature, we mainly focus on find the variables that affect the chronic absent rate and fit the model to measure the level of those independent variables' effects on absent rate. We use those variables to answer the following three questions: (1) What percentage of students are chronically absent in 2016? (2) How each school environment and characteristics variables influence the chronically absent rate? (3) How much of those independent variables impact the chronic absent rate?

In our analysis, we mainly focus on the 7 areas (Brooklyn, Bronx, New York, Staten Island, Jamaica, Flushing, Long Island City) with the greatest number of schools of New York City. The average percentage of chronically absent rate is around 23%. And this excessive value leads to society's concern. We find that the areas (i.e., New York, Bronx, and Brooklyn) with more schools lead to a greater absent rate. Based on the correlation table, we find that Percent Asian, Percent White, and Collaborative Teachers % is negatively correlated with absent rate. However, the Percent Black, Percent Hispanic, and Economic Need Index are positively correlated with absent rate.

Then, we add the population into the dataset. The regression plot indicates that the effects of population on absence rate are inapparent and there may exists endogeneity problem between Economic Need Index and Population. Thus, we need to cautious about this when we fit the model.

Later, we fit a regression model to find how much of those independent variables impact the chronic absent rate. In this part, we test the economic relationship between our independent variables and dependent variables. The scatter plots indicate a strong linear relationship between our Xs and Y. We recode the population variables into categorical variables to avoid endogeneity problem. The OLS

regression results show an increase in R-squared value as we add more variables to the model. In other words, a larger percentage of the variation in the absent rate is explained by our independent variables. After comparing 4 potential multiple regression models, we selected the one with largest R-squared value. The model is:

$$\begin{aligned} Model5: \widehat{ChronicAbsentRate_i} = {} & 0.5610 + 0.5618 EconomicNeedIndex_i - 0.7881 PercentAsian_i - 0.5601 PercentBlack_i \\ & - 0.6861 PercentHispanic_i - 0.4294 PercentWhite_i - 0.0187 pophigh_i \\ & - 0.1215 PercentCollaborativeTeachers_i \end{aligned}$$

All coefficients in this model are statistically significant since the P-value < 0.05. The characteristics variables (percent race and pop_high) and Percent Collaborative Teachers have a negative impact on Chronic Absent rates. The Economic Need Index is the only variable that has a positive effect on Chronic Absent Rate. It is hard to change the characteristic variables, so we want to minimize the adverse effects of environment variables (Economic Need Index and Percent Collaborative Teachers) on the absent rate. Based on the result I listed in the analysis part, to decrease the absent rate, the bureau of education should supervise more on the teachers' collaboration work and provide more economic support for those schools with a high economic need index.

There have several limitations to this model. The regression model results vary from the correlation table results because there are some unobservable effects on races and the absent rate. For instance, in our data, if we add the percent of each race in one school, the result is more than 100%. Some mixed-blood students may identify by two or more races that influence data accuracy. In future work, we may build another model to test those unobservable effects and use 2SLS method to deal with endogeneity problems.

We also have a large dataset that can support us apply the machine learning methods such as text analysis, regression trees, or random forests in future studies. We can use text analysis to research students’ attitudes towards the school environment from online postings. We can tackle both the

econometrics of the OLS and regression trees and the economic intuition behind both models.

If applicable, we can fit our model on other cities' dataset. The Economic Need Index and Collaborative Teacher Percent may or may not be the key issues for those cities' chronic absent rate. Adding other environment variables into the model will significantly improve our model's accuracy and explanation ability.

References:

1. U.S. Department of Education. (2019). Chronic absenteeism in the nation's schools: An unprecedented look at a hidden educational crisis. Accessed January 2019, at https://www2.ed.gov/datastory/chronicabsenteeism.html
2. Bartanen, Brendan. (2020). Principal Quality and Student Attendance. Educational Researcher. 49. 0013189X1989870. 10.3102/0013189X19898702. Accessed at https://www.researchgate.net/publication/338563399_Principal_Quality_and_Student_Attendance
3. Posterli, T. (2018, September 14). Did You Know That 28 Percent of New York Students Are Chronically Absent? What's The Solution? Retrieved December 18, 2020, from https://newyorkschooltalk.org/2018/09/know-28-percent-new-york-students-chronically-absent-whats-solution/
4. London, R.A., Sanchez, M., & Castrechini, S. (2016). The dynamics of chronic absence and student achievement. Education Policy Analysis Archives, 24(112). http://dx.doi.org/10.14507/epaa.24.2741