Introduction to Data Science, New York University, Summer of 2021

Author: Carlos Figueroa

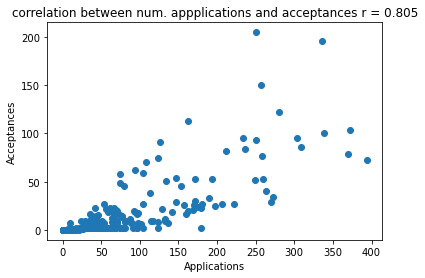
Title: Final project of the class

Dataset: The dataset (‘middleSchoolData.csv’) contains data from all 594 NYC middle schools, including 485 public schools and 109 charter schools (in the last 109 rows) from a randomly picked year in the past 5 years. Each row of the dataset represents a particular school, so the unit of analysis is “school”

Before he starts with out analysis, we must state the data-cleaning processes that were apply in this study to perform all of the analysis. The most important issue that most be addressed is that I am not taking into account any charter schools in my study, since there were some missing data points on it. Therefore, all the conclusions or assumptions that would be taken here are into the scope of public schools only, and it’s not entirely usable in the context of charter school. Moreover, I also skipped public school which had empty variables as well, and since this number was not too high, I consider that I wouldn’t affect our analysis in any sense. The final number of observations left in this dataset after the process just described was 449 observations (rows)

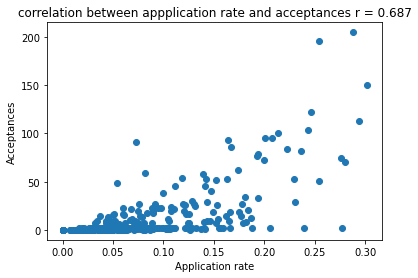
Question 1: What is the correlation between the number of applications and admissions to HSPHS?

Admissions stands for acceptances in our dataset. As found in this dataset, the correlation between these two variables is 0.805, representing a somewhat strong connection between both variables



Question 2: What is a better predictor of admission to HSPHS? Raw number of applications or application \*rate\*?

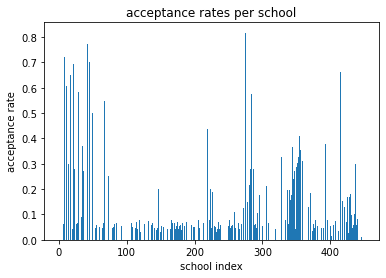
For application rate, I divided number of applications by school size of each school. To assess which one is a better predictor, we can first compare the correlations of each pair, and then square each correlation and compare which one accounts for more parts of the data. The higher the r-square, the better predictor a pair of variables is.



I found that the correlation between application rate and acceptances is 0.687, therefore its r-square is 0.4719. Whereas the correlation between number of applications and acceptances is 0.805, therefore its r-square is 0.6480. In other words, a model between number of applications and acceptances is a better predictor than by using application rate. And if we look at the data, after the 0.17 application rate in the x-axis, we can see that the reason why application rate is not a better predictor; it is because of multiple outliers that possess high levels of application rate, and lots of variability in their admissions (from zero acceptances to more than 200) and such variability is less when comparing raw application numbers (trending upwards). Moreover, we can also see that the school size impacts our analysis, and we can dig deeper into the reasons why afterwards.

Question 3: Which school has the best \*per student\* odds of sending someone to HSPHS?

For this analysis, I decided to create an individual acceptance rate from each school that applies to HSPHS. Basically, the higher their acceptance rate is, the more chances a school has of sending someone to HSPHS; in other words, due to the characteristics of such school, its students have more chances of being accepted if applying. To do this, I divided acceptance numbers by application numbers. Then, I got the highest acceptance rate from the dataset, and localized which school this rate belonged to.



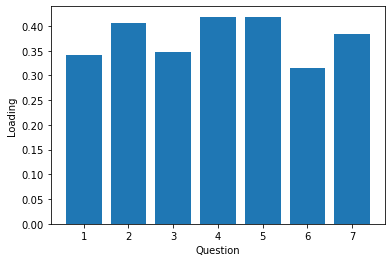
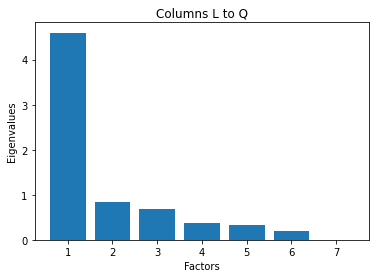
We can see that around the 270th school in our dataset, we had one of them with an acceptance rate higher than 0.8. I found that this school was “THE CHRISTA MCAULIFFE SCHOOL\I.S. 187” school, with an acceptance rate of 0.8167 or 81.76%

Moreover, for further analysis, I modified the acceptance rate calculations by not only taking into account the number of students that applied for HSPHS, but by using the number of students in the school. Hopefully, this modification would improve the conceptualization of “odds” from students that are applying and not applying (but could apply). With this modification, I found that “THE CHRISTA MCAULIFFE SCHOOL\I.S. 187” was still the best in terms of acceptance rate, now with such number as 0.234822 or 23.48%

Now, regarding another "odds" definition as from event happening divided by event not happening, I modify the acceptance rate by dividing number of students accepted into number of students accepted minus size of the school (which would account for all the students from that school that did not apply). Therefore, this calculation is no longer qualified as an acceptance rate, but as an odds probability. However, the best school with this calculation still was “THE CHRISTA MCAULIFFE SCHOOL\I.S. 187” with 0.306886 or 30.68%

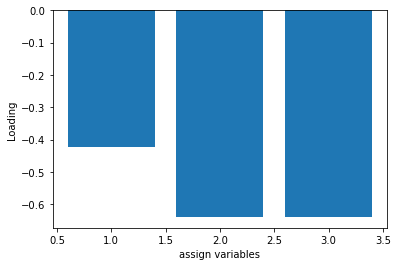
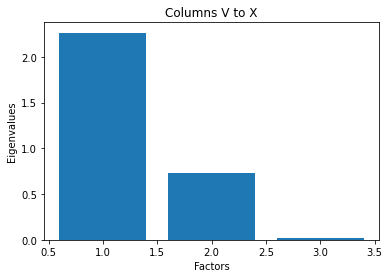
Question 4: Is there a relationship between how students perceive their school (as reported in columns L-Q) and how the school performs on objective measures of achievement (as noted in columns V-X).

For this question, I performed a PCA in how students perceive their school (columns L-Q) and another PCA for how the school performs on objective measures of achievement (columns V-X). Starting with how students perceive their school, we first check the eigenvalues:



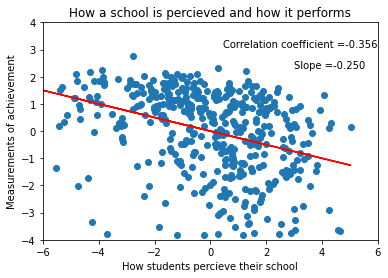
As we can see, by using the Kaiser line, we only have one meaningful factor with an eigenvalue higher than 1, which is factor 1. They all seem to be pointing at the same direction. Therefore, such principal component would be 1 dimensional (only one factor with eigenvalue higher than 1) and its name or representation (in order to account for where each question points at in our loadings matrix) could be “climate of the school” or “student perception of the school”

Now, for the second part, we have that when performing the PCA we obtain this:



A similar result, with one dimension being able to account for the other two factors as well. And when checking where does it points at. Moreover, here, we must remember in the interpretation of PCA, a negative loading simply means that a certain characteristic is lacking in a latent variable associated with the given principal component, but the direction where its pointing at is still interpretable. And to summarize this component, we will call it "Measurements of achievement"

Now that we found that each of these subsets of column can be represented in one dimension, we can proceed to create a scatterplot by using both variables and fit a linear model.



This model has an r-squared equal to 0.127 and a p-value of 0.0000000. Therefore, it is somewhat statistically significant, but does not cover for most of the data. And now, when discussing is correlation coefficient, we see that it is negative (meaning that when students perceive their school as pretty good, the measurements of achievement of that school tend to go down) and is not that high in order to say that there is a strong relationship (which can also be argued by considering that it was such a low r-square). In other words, there seems to be a negative relationship between how a school is perceived and how it performs in objective measurements, but such relationship is somewhat weak to generalize it for all public schools in our dataset. But by looking at the scatterplot, we can see that by ignoring the clusters in the middle and the variance in the tails, there seem to be a considerable number of cases where schools perceive their climate as pretty good to almost perfect and perform horribly in objective measurements. But outliers such as schools with good climates and good measurements weakens such relationship, which is something that we could have expect from the beginning, exceptional schools.

Question 5: Test a hypothesis of your choice as to which kind of school (e.g. small schools vs. large schools or charter schools vs. not (or any other classification, such as rich vs. poor school)) performs differently than another kind either on some dependent measure, e.g. objective measures of achievement or admission to HSPHS (pick one).

To answer this question, my hypothesis is that rich schools, or at least schools in which students average income is higher than the median student average income from all schools, tend to perform better in objective measures of achievement than poor schools, or schools bellow the median student average income from all schools. To do this, I create two categorical data columns; the first one stores 1’s if average income is higher than the median average income from all schools, and 0’s if its lower, and the second one translates 1’s into “high income” and 0’s into “low income”

With the uncertainty of the distributions of these variables, and with the possibility of also categorizing schools’ rates of achievement into high achievers and low achiever, we can run a non-parametric test. First, we will say that the scale from which our variable objective measurements of achievement ranges between -6 and 6, and if a school’s rate of achievement is higher than 0, we will categorize them as high achievers, and if lower than 3 as low achievers. Now, we create a cross table of our categorical variables and we obtain this:

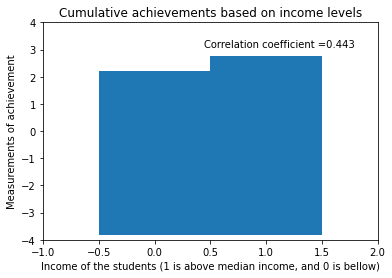
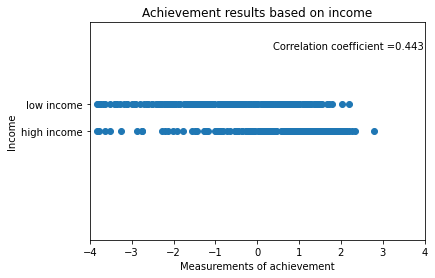
high income low income

high achieve 176 82

low achieve 48 143

From this point, we can see that most of the schools that have high levels of achievement are schools that are above the median income average from all schools, but we must first test this to see how reliable it is. So, our null hypothesis would be that high levels of income do not has an effect in achievement scores, and our testable hypothesis is that it does. For this case, we will perform a chi-square test. I obtained 79.7832044952326 as the chi-square value for the cross table presented before, with a p-value equal to 4.178294361212231e-19. If we consider our critical p at 0.01, we will have to reject the null hypothesis, and denote that there could be a meaningful relationship between income and achievement scores, but such hypothesis cannot be proven, only considered.

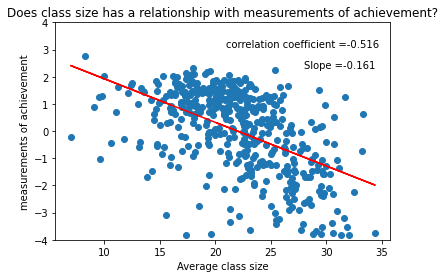
Now, with one categorical variable (high income or low-income schools) and a continuous variable (objective measurements of achievement that we encounter into 1 dimension in question 4) we could also show it in a scatter plot to see if we note any differences:

For this analysis, we applied a point-biserial correlation coefficient which uses a binary variable and a continuos variable, and we found that such correlation is 0443 with a pvalue less than 0.05. Moreover, in the bar char we can see that there is higher numbers of cumulative achievements in students with “high income” and in the scatterplot at the left, we can see that there are more points in which achievement measurements are higher than zero in the line of high income than in the line of low income, and the same thing is true when talking about low income schools and low measures of achievement.

Question 6: Is there any evidence that the availability of material resources (e.g. per student spending or class size) impacts objective measures of achievement or admission to HSPHS?

I decided to test if there is any evidence that the class size of the school impacts objective measures of achievement. To do so, I used the average class size of each public school and the objective measures of achievement from question 4. Classic literature dictates that the smaller the class size is, the better professors can impact on the learning of the students, since they don’t have to focus on a wider range.



For this model, we have a correlation of -0.516, r-square of 0.266 and a p-value less than 0.05. We can see that the model seems to predict the trend from the majority of datapoints, but there are some outliers that reduce our correlation coefficient and therefore our r square. The existence of these outliers could be because some schools that perform badly on measurements of achievement have small class sizes, but the reason why they are so low in the measurements is because of other conditions like income or ethnicity. In other words, we should create a multivariable model that accounts for all the different factors that could potentially affect the measurements of achievement in public schools. However, by adding more factors, its not going to potentially increase our correlation coefficient, and we could agree that average class size is a pretty good predictor in this context.

Notwithstanding, there is plausible confounders to this relationship that we could control, and its income. Students from rich schools can afford outside-school instructors and schools like that tend to attract better professors to their schools. Therefore, we could perform a partial correlation, accounting for average income from the schools, and seeing how strong the relationship between average class size and objective measurements is. In this partial correlation, I found these outputs:

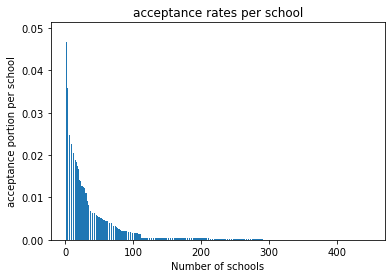
n r CI95% p-val

pearson: 449 -0.386 [-0.46, -0.3] 0.0

Which means that the partial correlation between class size and measurements of achievement is -0.386 while controlling for income levels (in a non-categorical way). Our conclusion is that class size does have an effect in objective measurements, not very strong, and improvements in our model (by adding more variables or controlling more confounders) could potentially control and improve this model.

Question 7: What proportion of school’s accounts for 90% of all students accepted to HSPHS?

To access this question, I first took the sum of all the students from public schools accepted into HSPHS, which is 4192 in this dataset (excluding charter schools). Then, I created a column in which I stored the portion each school has in the total population of students accepted into HSPHS, by dividing each acceptance column by 4192 and storing their portion contributed. Then, I created a column that stored the cumulative sum of all these portions of acceptances from each school, and then I sorted these values from lowest to highest in terms of the percentage each school represents on the whole population of students accepted and counted the indexes (after checking there were no empty spots due to data cleaning methods) until the cumulative column was equal to 0.90 or 90%. What I found is that 92 schools (starting from the ones that have the biggest contribution to the whole population of students accepted into HSPHS) account for 90% of all the students admitted. In other words, 20.49% of all the schools accounted for 90% of all students admitted into HSPHS with this metric. Now, if we start from the schools that have the lowest contribution to the top, and we exclude the schools that have the biggest portions, we will have that 447 account for 90% of all students accepted to HSPHS, in other words, 99.5% of the schools account for 90% of all students accepted by this order from lower to higher. This also means that 0.5% of schools not accounted by the 99.5% represent the resting 10% of students admitted, giving them the title of having the biggest impact into the proportion of students accepted into HSPHS.



In this graph we can see that there is a school that almost has 0.05 impact in the whole population of students being accepted into HSPHS. This graph could help clarify the two methodologies used before. The first one goes summing from left to right and stopping when it reaches 90%. The second one goes from the right to the left, and stops when it reaches 90% as well, but because of the skewness, both methods throw different percentages of total schools accounting for 90%.

Question 8: Build a model of your choice – clustering, classification or prediction – that includes all factors – as to what school characteristics are most important in terms of a) sending students to HSPHS, b) achieving high scores on objective measures of achievement?

To assess question A in terms of sending students to HSPHS, I constructed a multiple linear regression as a prediction model, and used all the available variables to see which one has the biggest weight when talking about increasing the number of acceptances into HSPHS. The weights of each variable were:

'applications' = 3.29964306e-01 'per\_pupil\_spending' = -2.55093905e-04

‘rigorous\_instruction' = -5.92077563e-01 ‘collaborative\_teachers' = 2.16282602e+00

‘supportive\_environment’ = 2.10380258e+00 ‘effective\_school\_leadership' = -2.64641847e+00

‘strong\_family\_community\_ties' = 3.85667884e+00 'trust' = -3.44781314e+00

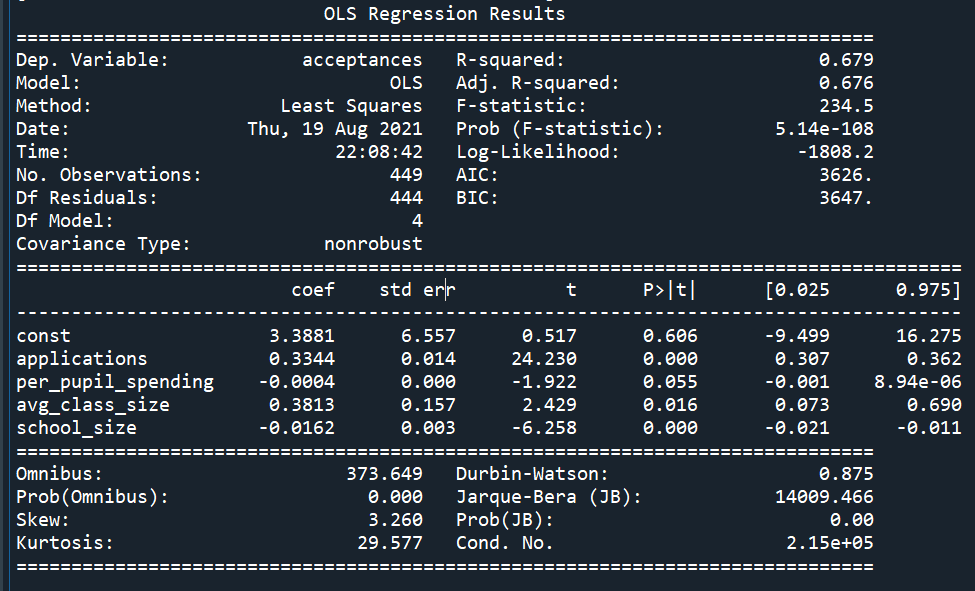
‘school\_size' = -1.37191072e-02 'avg\_class\_size' = 2.75614196e-01

'objectivePerformance' = -1.18724230e+00 (obtained from PCA of columns v to x)

'climate' = -3.26640936e-01 (obtained from PCA of columns l to q)

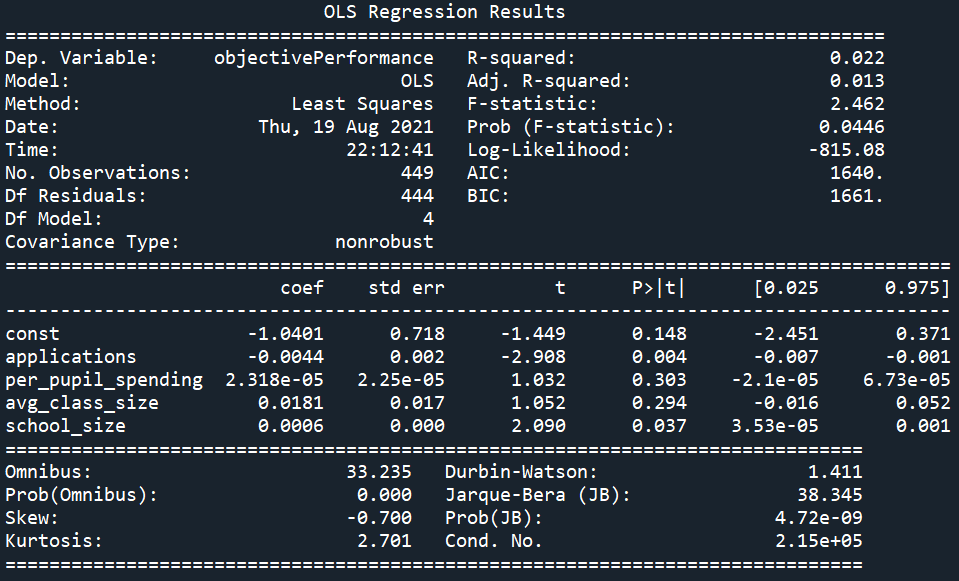
The drawback of this analysis is that it is not using categorical variables, and its only giving us the change in acceptances whenever one of these variables increases by one. As we can see, the variables with the strongest weights or coefficients are variables that don’t possess too much variance, in other words, discrete variables. And the problem with these variables is that they don’t change too much and become misleading when trying to predict. Therefore, we should not consider them for our analysis, but we can see that they are good predictors in the increase of acceptances in a small way (most of them go from 0 to 5, so in the best-case scenario, an increase in strong family and community ties could improve acceptances of the school by 19).

Now, focusing on the continuous variables left, and construction another multiple variable model with them, we have that:



First, we can see that all of our results are statistically significant (p< 0.05). So, by controlling for confounders with our model, and only using these variables, we can finally say that the number of applications is the biggest predictors of higher acceptances in a school. However, this fact is almost intuited, the more people that apply, the more acceptances the school gets. Hence, we should focus on the second heaviest factor, which is average class size (a continuous variable which range is not that big) appearing to have a 0.3812 change in acceptances by a one unit increase in it. In conclusion from this analysis, average class size is the most important factor found in our dataset to send a student to HSPHS. However, its coefficient is not that high, so there is not much reliability included.

Now, to assess question B, and knowing what we know from our past analysis, we will use the last 4 variables used, and try to predict objective achievement measurements.



From this, we can see that average class size is the most important variable in this model, since it has the highest coefficient. However, its p-value indicates that it is not statistically significant, but still somewhat important in obtaining better results in objective measurements of achievement.

Additionally, we could perform further analysis by creating dummy variables in categorical variables such as income level. For instance, by using the data column we curated in question 5, the level of income could be measured by tagging them as rich if the ‘per\_pupil\_spending’ of each school is higher than the median value of this variable across all schools. Then, if rich is 1, poor is 0, and see how much such variables account for acceptances. The same principle could be applied to the variables described as discrete paragraphs ago (above or below 2.5 could represent the existence or scarcity of them) and the percentages of races in the school (0.50 stablish a majority of population with an ethnicity). However, for the purpose of this study, I did not cover such topics.

Question 9: Write an overall summary of your findings – what school characteristics seem to be most relevant in determining acceptance of their students to HSPHS?

Applying all the techniques I have learned throughout the class with a real dataset was pretty exciting, but there were meaningful time constrains and some efforts I could not do due to that. This dataset turned out to be more interesting than expected, and somewhat tricky to work with in some parts. Regarding my findings, I will say while responding to the first questions, the analysis was fairly simple (creating some application rates and checking which school was the best in terms of sending its students to HSPHS) and towards the middle I started to focus on the impact different variables had in objective measures of achievement (question 4, 5 and 6). I found that class size and income seem to be the most important factor impacting these measurements, and I dig deeper into them on question 8, finding that the most important between both was the average class size. However, combining these special characteristics work pretty well while constructing a predictive model from this dataset; for instance, the last model created on question 8, part A had an r-square of 0.679.

Question 10: Imagine that you are working for the New York City Department of Education as a data scientist (like one of my former students). What actionable recommendations would you make on how to improve schools so that they a) send more students to HSPHS and b) improve objective measures or achievement.

Based on the findings I got from this study, and on the things that we can change, I will say that the New York City Department of Education should focus on the average class size of their schools in order to improve sending more students to HSPHS and improving their objective measurements of achievement, as discussed in question 6 and 8. Moreover, they should also focus on promoting students to apply for HSPHS, since it seems like not many students from average schools are applying, and that creates cases where some schools do not even have one accepted student into HSPHS.

PD: If there are any doubts regarding the code I used, please do not hesitate to email me for it. I will incluse a link to a google sheet with all of it here:

https://docs.google.com/document/d/1fvqbOcWNsi3XgSrLiIdmfkXmnDwAFpM7EZ0bZf5DGg4/edit?usp=sharing