Predicting College Enrollment Rate in CPS High Schools

Erik Bergmark

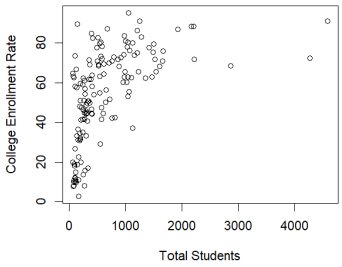
The Problem

Many public high schools set out to prepare students for further education. Chicago Public Schools (CPS) high schools are no exception. In fact, CPS students enroll in college at a higher rate than the rest of the country. In 2024, 65.2% of CPS students enroll in colleges in the fall after graduating. This rate is 62.8% on the national level (U.S. Bureau of Labor Statistics, 2021). However, not every public high school in Chicago has an enrollment rate this high. This report aims to explain where the discrepancies in enrollment rates within Chicago might come from, and to identify what features of high schools are associated with changes in enrollment rates.

The Data

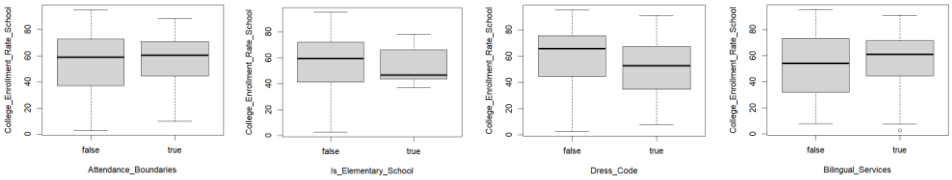
The data used for this report was found from the Chicago Data Portal. It is a dataset of school profile information about all public schools in Chicago for the 2024-2025 school year. There are 652 rows and 99 columns, where each row represents an individual public school. For the purposes of this analysis, we will reduce the dataset to specific columns that are relevant to us and filter the rows to just high schools. We are left with the following variables: Whether or not the school is also a middle school, and/or an elementary school. The title of the school's administrator, either principal or director. Whether or not the school has attendance boundaries. The total student count, as well as counts for specific demographics, including white, black, Hispanic, Asian, Native American, Asian-Pacific Islander, Hawaiian-Pacific Islander, or other. Whether or not the school has a dress code. Whether or not the school has bilingual services. Whether or not the school is a GoCPS participant. And of course, college enrollment rate, which is our target variable.

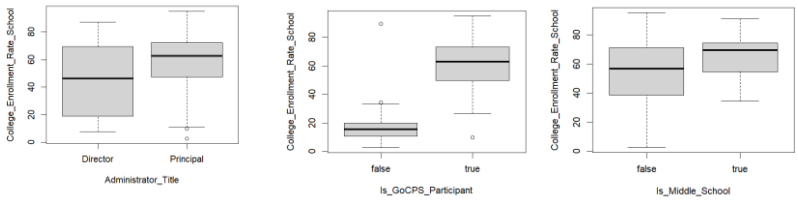
Preliminary Analysis



Before we begin modeling, we will try to gain a prior understanding of how these variables interact with the target variable, and whether we can expect relationships.

First, we are looking at college enrollment rate plotted with the total number of students in the school. There is some relationship between the two, but not a very linear one. There are two schools with over 4000 students and relatively high graduation rates which could be highly influential points. We will check these points later.



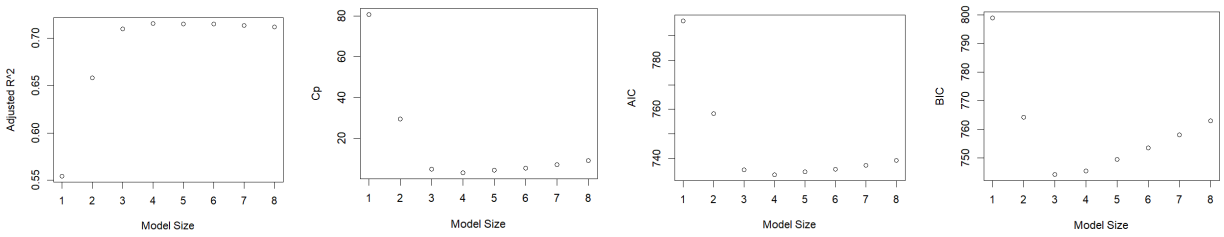


The boxplots above display the difference in the distribution of college enrollment rate for our categorical features. Attendance boundaries seem to have no impact on enrollment rate. A school with an attendance boundary allows students in the neighborhood automatic enrollment. The most significant difference in groups is whether the school is a GoCPS participant or not. GoCPS is an admissions platform. It is often used by schools with selective enrollment, or advanced programs like IB. It makes sense that these kinds of schools have higher college enrollment rates.

Modeling

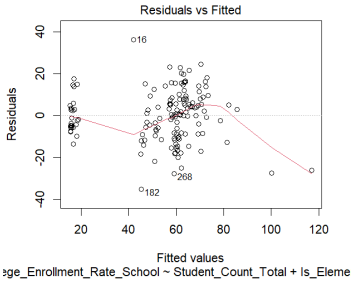
Our goal is to build an interpretable linear regression model that best captures the relationship between the given features and the response, which is college enrollment rate in this case. The first step is to fit the full model with every feature, to get a basic understanding of the linear relationships. This model has a decent adjusted r-squared value of 0.6728, but there is a glaring issue with this model. When we look at the variance inflation factors of each predictor, several are extremely high, much higher than our rule of thumb of 10. These are the student count variables. This is intuitive, schools with higher student counts in total will have higher counts for each specific demographic. Two ways to approach this issue are principal component analysis and ridge, but because one of our goals is to maintain interpretability, these are not the best solutions. Since we are dealing with near perfect multicollinearity, these count variables are likely redundant. So, we will opt to just keep the total student count.

The next step is to determine which of these uncorrelated variables to keep. We don't want to keep variables that don't have enough predictive power, to maintain a tradeoff between variance and bias. We will use the leaps package in R to find the best model for each possible number of predictors and compare model evaluation criteria.



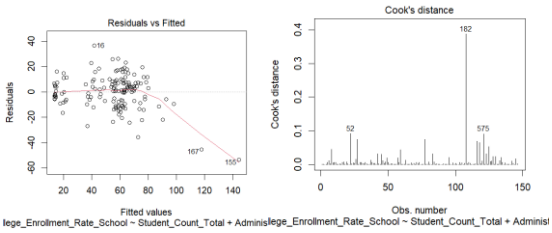
We are looking to maximize adjusted r-squared, and minimize Mallow's Cp, AIC, and BIC. All criteria select 4 variables, except for BIC which selects 3. We will choose 4. These variables are the total student count, whether the school is also an elementary school or not, whether the school has attendance boundaries or not, and whether the school is a GoCPS participant. Our next step is to fit the model with these variables. This model has an adjusted r-squared of 0.7155, which is higher than our initial model. However, we run into a new

issue. When we plot the fitted values against the residuals, we see evidence of heteroscedasticity. The Breusch-Pagan test agrees with this evaluation, with a p-value of 0.006726.



It makes sense that we see this issue. College enrollment rate is found to be the total number of students enrolled divided by the total number of students, and we are using number of students as a predictor. The variance will be a function of the number of students. To fix this, we will perform variable selection with again, but with weights of 1/total student count. While 4 variables are still selected, the administrator title variable is selected instead of the elementary school variable. The difference between a principal and a director administrator is not explicitly defined, but we can imagine that a director is likely a more non-traditional role.

As shown in the new residuals vs fitted plot, our heteroscedasticity issues have been resolved. The Breusch-Pagan test now has a p-value of 0.9996. With this model, we have no issues with multicollinearity, and looking at Cook's distance, we find no highly influential points above our threshold of 1. This final model has an adjusted r-squared value of 0.7998, which is our highest yet. 79.98% of the variance in college enrollment rate can be explained by the four variables we have selected.



When we hold all other variables constant, here is how we can interpret the coefficients of our model: For each additional student, we expect college enrollment rate to increase by 0.019487 percent. The expected college enrollment rate for schools with a principal administrator as opposed to a director is 4.111766 percent higher. The expected college enrollment rate for schools with attendance boundaries as opposed to without is 20.604148 percent lower. The expected college enrollment rate for schools that participate in GoCPS is 38.897104 percent higher than schools that don't.

While correlation does not equate to causation, this model can help schools understand what might be helping or hurting college enrollment rates. Of course, there are outside interactions that this model does not consider. Even though the coefficient for total students is positive, schools should not enroll students that they do not have the resources to support just to try to improve college enrollment. This may have the reverse effect. The model also tells schools what might not be important in predicting college enrollment. For example, since the dress code variable was left out of the model, schools do not have to worry about enforcing a dress code or removing their dress code with the hopes of improving college enrollment.

This model is a good start, but to be as effective as possible it needs to consider more features. For example, socio-economic statistics about the neighborhood each school is in could be good indicators of college enrollment rates.

Hopefully Chicago Public Schools can continue to support college enrollment above the national average and try to improve rates by understanding what schools with high enrollment rates are doing right.

Sources

Chicago Public Schools. (2024, November 10). *Chicago Public Schools – School Profile Information SY2425*. https://data.cityofchicago.org/Education/Chicago-Public-Schools-School-Profile-Information-/3dhs-m3w4/about_data

U.S. Bureau of Labor Statistics. (2021, April 27). *College Enrollment and Work Activity of Recent High School and College Graduates Summary*. Bls.gov. https://www.bls.gov/news.release/hsgec.nr0.htm