Lindsey Dorson

Data Origins and Background

All the data used in this project was collected from the Governor's Office of Student Achievement (GOSA) website. The GOSA is a government organization in the state of Georgia that is dedicated to improving the state's education within all education levels (pre K to college). They regularly update publicly accessible county and school level data regarding different educational factors and outcomes. This data is collected from individual school systems within the state and is updated regularly throughout the year. Information on class sizes, graduation outcomes, and standardized testing are examples of what is collected by the GOSA. This information is made publicly accessible for researching purposes and for transparency among concerned parties on the different education systems in Georgia. It is important to note that only data from as far back as 2010 are listed on the government website. However, older data can be requested if needed.

The main dataset used in this project was the Post-Secondary C11 Report. This dataset gives information on the number of students that are enrolled in a post-secondary institution within 16 months of graduation. These values are aggregated by both school county and individual high school. Within those specifications, demographic information on the student graduates is also present. This is nice because these demographics variables can give insight on some of the external factors that might affect college enrollment. Initially, this dataset was collected as cross-sectional data from the many downloadable csv and excel spreadsheets that are accessible on the website and separated by individual years (later converted to panel). There is usually no codebook to go along with the datasets, however, the variable names are sufficient in indicating the variable values and meanings. Also, there is a small informative paragraph usually listed next to the spreadsheets available for download on the website.

For my project, I specifically downloaded this dataset to represent the school years 2015, 2016, 2017, and 2018 (4 spreadsheets in total). All these individual datasets combined made up 2,532 observations with each year containing around 630 observations by itself. There were also 33 descriptive variables that as mentioned earlier gave quantitative information regarding the number of graduates and future college enrollments from those graduates on both an individual school or county level. All the counties and high schools in Georgia were represented in this data. This makes it possible to research what factors contribute to the higher percentage of student graduates that continue their education after high school. This will be the purpose of my project, but my unit of observation will be individual high schools in Georgia. I will also include other variables that are gathered from additional datasets also from the GOSA websites. These variables will come from the Direct Certification (School Level), Hope Eligible Graduates, and Enrollment by Grade Level datasets. Each of these will also be collected for each of the years 2015-2018.

Describe the Raw Data

The raw data contains variables that indicate for each observation the year of observation, county codes, school codes, school names, number of total graduates, number of total postsecondary enrollment 16 months later, and number of graduates and postsecondary enrollment (for each subgroup). All the numerical data values are whole numbers and the important subgroups that are also included in the data are males, females, free reduced lunch students, white students, and black students.

Importing the raw Post-Secondary C11 Report data (years 2015-2018) from an excel file into R proved to exhibit some challenges. All the numeric variables were not properly formatted and appeared as string values. Also, the variable names for the different subgroups were untidy and long (Ex. "Total High School Graduates…8" and "Number of High School Graduates Enrolled in Postsecondary Institution…9"). Therefore, all the variable names need to be renamed for easier interpretation and shortened. Also, almost all the variable names contained multiple spaces within itself.

There were 2,532 observations across all years in the data set (when combined). However, when not considering the aggregate values for each county, there were 440 high schools listed in the year 2015, 439 high schools listed in the year 2016, 443 high schools listed in the year 2017, and 453 high schools listed in the year 2018. This means there was not a huge variation in the number of high schools each year.

A positive is that there were no missing values present in the dataset. However, there were a significant number of values labeled "TFS" which the original excel spreadsheet indicated represents "too few students". In fact, "TFS" makes up 36,007 values out of the 83,556 total values listed in the raw data which amounts to roughly 43% (all years combined). Each year had around 9,000 values listed as "TFS". Also, "TFS" is also mainly present in the columns for smaller subgroups like migrant students, limited English proficient students, American Indian or Alaskan Native students, and Pacific Islander students. As expected for variables representing larger subgroups like male students, female students, white students, and black students the value is far less frequent. Smaller high schools also had this value frequently pop up as well. Upon further inspection "TFS" indicates values between 0-9 so the value can be interpreted with a range of numbers.

While looking at the variables total high school graduates versus total postsecondary enrollment (16 months later) in a scatter plot a few outliers were presented. However, upon inspection these outliers were just the aggregate county total of all graduates and postsecondary enrollment students for each individual year. When I plot the total graduates versus total postsecondary enrollment with only individual high schools, there are no outliers that are technically present. This is because with schools that are larger and as a result have a higher number of graduates there is a higher chance of postsecondary enrollment. Therefore, it makes sense there would be some plots that show up far from the main cluster because these plots represent the larger high schools.

Getting an Analytic Dataset

To get the analytic dataset multiple steps need to be conducted. The first step is combining/joining the individual year datasets into one large data frame. Next, aggregate county amounts were filtered out so only individual high schools were left across the years 2015, 2016, 2017, and 2019. The next and by far lengthiest step is renaming all the variables to have shorter/easier names and no spaces between their names. This is repeated for all variables and is especially helpful for interpreting and cleaning the many subgroup variables. Now variable names that were previously "Total High School Graduates…8" and "Number of High School Graduates Enrolled in Postsecondary Institution…9" can be simplified to better indicate their appropriate subgroup which in this case is male graduates. This is repeated for all the subgroups mentioned previously. Then string columns for the quantitative variables were converted to numeric columns. This transformation led to all "TFS" values turning into missing values (NA).

The missing values appeared significantly in smaller subgroups as mentioned before. Therefore, complementary variables (number of graduates & number of postsecondary enrollment) for each subgroup that contained higher than 29% missing values were dropped from the analytic dataset. This left the number of graduates and the number of post-secondary enrollment variables for male graduates, female graduates, free reduced lunch graduates, white graduates, and black graduates.

The school code variable column had differing character counts for each year and also in future merged data sets. For example, Alexander High School in the years 2015, 2016, and 2017 had a school code of “187” while in the year 2018 it had a school code of “0187”. To correct this, a “0” was pasted in front of school codes that had three characters.

The last step in cleaning the data set was creating a balanced panel. In other words, high schools that were included in all 4 years of observation were kept for the analytic data set. This was done by first creating a new variable for group id that was composed from concatenating the county code with the corrected school code for each high school. Then giving each group id a unique value for which the frequency of occurrence was counted (new variable). If the count of occurrence was less than 4, then the high school was dropped.

Merging Other Data

Now that my dataset is all clean, I merged in some new variables from datasets that were also collected from the GOSA website. These datasets were called Direct Certification (School Level), Hope Eligible Graduates, and Enrollment by Grade Level (all originally cross sectional by year). As I did with the main College Enrollment dataset, I only collected each one from the years 2015-2018. These datasets were added to control for more factors that could affect a student’s decision to enter a postsecondary institution. Also, all of them were cleaned like the initial dataset but there were extra steps taken for the Enrollment by Grade Level dataset which will be specified below.

Financial support though scholarships could play a large part in a student's decision so that is why I added a variable indicating the number of graduates that were eligible for the HOPE scholarship and a variable indicating the percentage of hope eligible graduates by school level. This was from the Hope Eligible Graduates datasets. Also, a variable indicating the percentage of students that were directly certified was added from the Direct Certification datasets. It is important to note that for a student to be considered "directly certified" the student must have been a foster child, homeless, and or received government assistance. The last variable merged in was from Enrollment by Grade Level datasets. This dataset had a few more columns compared to the other ones and included all schools in Georgia that served any grade level. Therefore, the schools/observations that served grades other than 9th,10th, 11th, and 12th were dropped. Also, only observations for the enrollment period of Fall and that contained the grade level 12th were kept as well. This leaves only the traditional high school's enrollment count in the Fall for seniors in Georgia. This initial enrollment variable was the last variable merged in.

All the datasets were merged in one by one to the initial dataset by the columns: year, school name, county name, county code, and school code (important identification columns). Schools that did not match the initial data set were not included after the merge. Also, the same steps to make a balanced panel was redone to correct for any duplicate high schools resulting in merge. This leaves 1,312 observations in total across the 4 years of observation. In other words, there are 328 high schools observed each year.

Table 1.

Sample Size by Year

	2015	2016	2017	2018
Number of Observations	328	328	328	328

Note: Panel is balanced by year

Imputation

Since I have all datasets merged, I next fixed the missing values that resulted from the "TFS" values being converted to numeric. These missing values were imputed with a numeric value of 5. 5 was selected because it is the rounded whole number resulting from the median of the range 0-9 and as mentioned earlier "TFS" is between 0-9.

New Variables Created

1) "Highwhites": dummy variable with 1 indicating more than 50% of a schools graduates were white.

2) "HighFRL": dummy variable with 1 indicating more than 50% of a school's graduates received free reduced lunches.

3) "Collenroll_per": percentage of total graduates that were enrolled in a postsecondary institution (two decimal places).

4) "Totgrads_per": percentage of total graduates compared to initial senior year enrollment at high school (two decimal places).

5) "HighHope": dummy variable with 1 indicating more than 50% of school's graduates were eligible for HOPE scholarship.

6) "HighCert": dummy variable with 1 indicating more than 50% of school's graduates were directly certified.

7) "Cenrollbase_per": base percentage of total college enrollment compared to total high school graduates averaged from all years (two decimal places).

8) "Cenrollbase_meets": dummy variable with 1 indicating a school's college enrollment percentage was at or above the base percentage from before.

Exploratory Analysis

I am mainly interested in how the HOPE scholarship affects college enrollment in Georgia. I predict that the HOPE scholarship has a positive effect on college enrollment. This is because it provides a reduction in the financial burden of attending college. In a way it can incentive more Georgia students to receive a college education.

My final analytic data set included quantitative variables for the total number of high school graduates, the total college enrollment in postsecondary institutions, and all the new variables I created in the previous section for each unit of observation across years 2015-2018. The subgroup variables were not included because they required too much imputation which could affect my analysis. However, my analytic dataset's quantitative variables now mainly include dummy and percentage variables. In final form, the dataset is also considered a balanced panel data set which was initially compiled from many cross-sectional spreadsheets.

Figure 1.

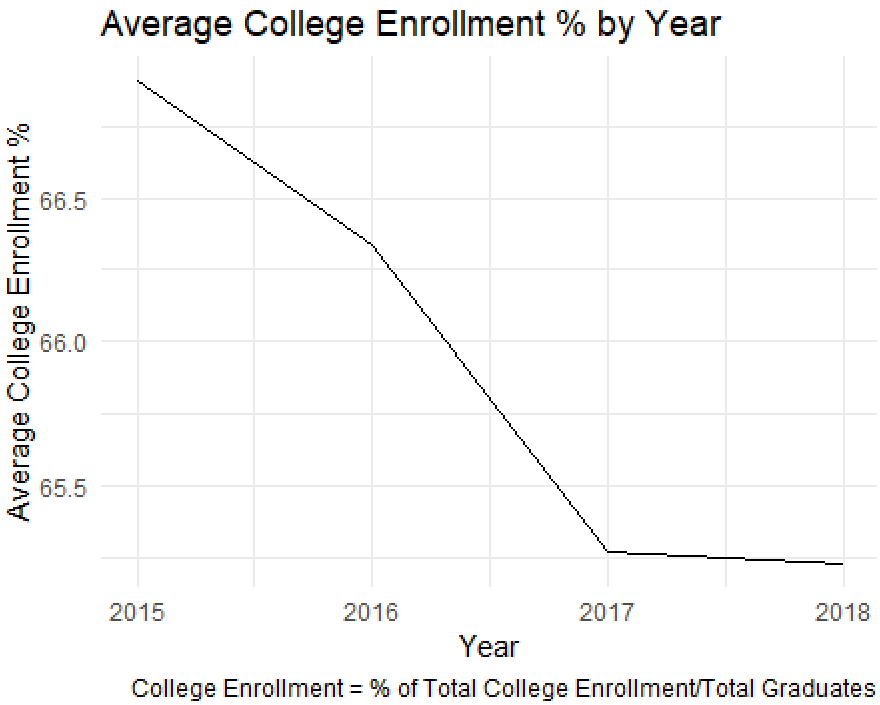

Figure 1 illustrates the average college enrollment percentage by year. While there appears to have been a reduction in the average percentage the difference is minimal judging by the y axis values.

Figure 2.

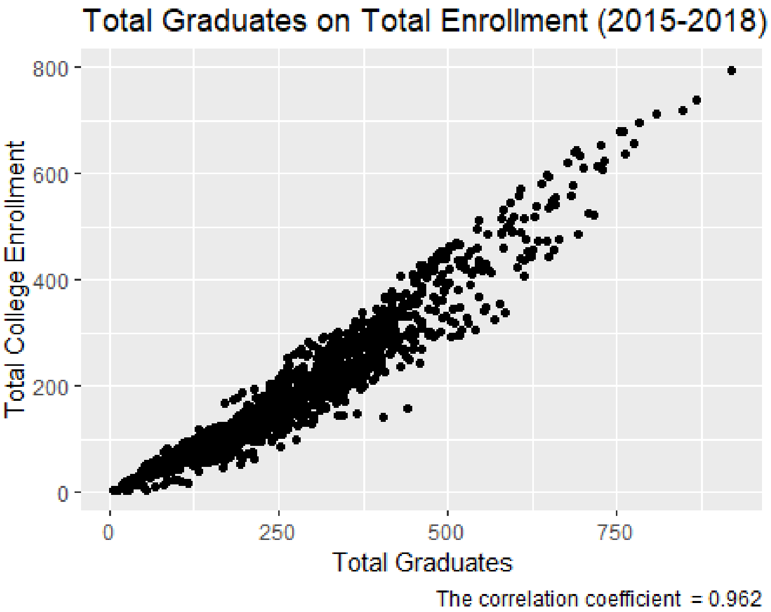

Figure 2 plots the number of total college enrollment by the number of total graduates for all years. The linear relationship appears very strong. The exact correlation coefficient between total number of graduates and the total college enrollment is .962. This means that the total number of graduates and the total college enrollment have a strong, positive linear relationship (as the figure illustrates). In other words, as the total number of graduates in each high school increased so did the total college enrollment across the years of observation. However, this is not surprising considering a larger number of graduates gives way for a potentially larger number of graduates that can be enrolled in college.

Figure 3.

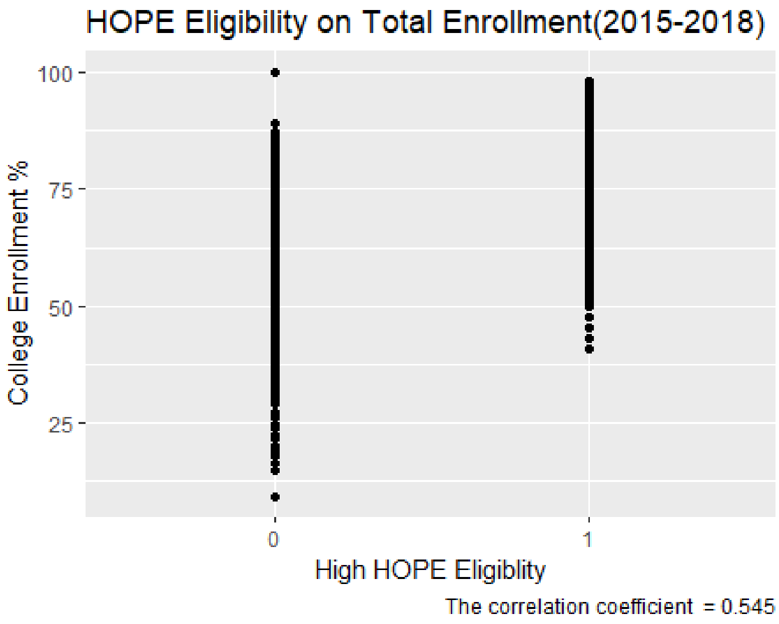

Figure 3 plots the college enrollment percentage for observations across all years and is grouped by high HOPE eligibility. It appears that high HOPE eligibility is correlated with higher percentages of college enrollment. The exact correlation coefficient between the two variables is .55. This means that there is a positive, linear relationship between a high school's college enrollment percent and if it had high HOPE eligibility. However, this doesn't give a causal interpretation for the relationship between the two variables.

Table 2.

Summary Statistics

Variable	**N**	**Mean**	**Std. Dev.**	**Min**	**Pctl. 25**	**Pctl. 75**	**Max**
TotGrads	1312	280.557	152.144	5	168.75	373	920
TotCEnroll	1312	195.088	131.673	5	98	265.25	793
Highwhites	1312	0.495	0.5	0	0	1	1
HighFRL	1312	0.534	0.499	0	0	1	1
Collenroll_per	1312	65.936	13.794	9.26	57.408	75.452	100
Totgrads_per	1312	97.758	13.328	2.94	93.295	100.252	362.5
HighHOPE	1312	0.322	0.468	0	0	1	1
HighCert	1312	0.083	0.276	0	0	0	1
Cenrollbase_per	1312	65.94	0	65.94	65.94	65.94	65.94
Cenrollbase_meets	1312	0.486	0.5	0	0	1	1

Table 2 shows the summary statistics for my main variables of interest and included all years of observation. In my sample, the average number of total graduates was 281 graduates while the average number of graduates later enrolled in a post-secondary institution was 196. However, this doesn’t factor in the varying high school sizes. An average of 50% of schools had a high proportion of white graduates compared to other races. An average of 53% of schools had a high proportion of free reduced lunch recipients. An average of 8% of schools had a high proportion of directly certified graduates. Surprisingly, an average of only 32% of schools had a high proportion of HOPE eligibility. Another interesting observation is that only 49% of observations met or exceeded the base college enrollment percentage (65.94 %) which was averaged from all years.

Regression Models

To find the causal effect of the HOPE scholarship on college enrollment two fixed affects regression models were used and controlled by each year. Both were fixed by individual school id, however, the first one had no controls and the second one added controls in. The main dependent variable in these models were college enrollment percent and the main explanatory variable was the dummy variable for high HOPE eligibility. Table 3 shows the model results. The main explanatory variable maintained its significance with both models and the coefficient only reduced by a small amount when controls were added. This means HOPE eligibility is significant in predicting college enrollment percent. Also, its positive coefficient means it positively affects future college enrollment percent as well. Another interesting result is that schools that had a higher proportion of directly certified graduates also seemingly had higher college enrollment percentages. I would have expected the opposite result. However, the variable shows high significance.

Table 3.

Model Summary

Dependent Variable: Collenroll_per	**Naive**	**Controls**
HighHOPE	3.1010***	1.8549***
	(0.6003)	(0.5368)
as.factor(Year)2016	-0.5441	-0.5346
	(0.3418)	(0.3018)
as.factor(Year)2017	-1.5879***	-1.1365***
	(0.3418)	(0.3031)
as.factor(Year)2018	-1.7037***	-1.2975***
	(0.3418)	(0.3034)
Highwhites		0.2465
		(0.8286)
HighCert		4.2774***
		(1.2388)
HighFRL		-0.8267
		(0.5560)
Cenrollbase_meets		7.6553***
		(0.4738)
Observations	1312	1312
R2	0.059	0.270
R2 Adj.	-0.259	0.020

* p < 0.05, ** p < 0.01, *** p < 0.001

Note: Standard Errors in Parenthesis

Date Reference

The Governor's Office of Student Achievement. (n.d.). Retrieved from https://gosa.georgia.gov/