2021

Regression Analysis: Graduation Rate in Kentucky Public High Schools

Rebecca Price
Western Kentucky University, rebecca.price613@topper.wku.edu

REGRESSION ANALYSIS: GRADUATION RATE IN KENTUCKY PUBLIC HIGH SCHOOLS

A Capstone Experience/Thesis Project Presented in Partial Fulfilment

of the Requirements for the Degree Bachelor of Science

with Mahurin Honors College Graduate Distinction

at Western Kentucky University

By

Rebecca Price

May 2021

CE/T Committee:

Dr. Nicholas Fortune, Chair

Dr. Thomas Richmond

Dr. Melanie Autin

ABSTRACT

Kentucky's Public High School graduation rates vary widely across the rural and urban regions in the state. In addition to their graduation rates, each of these schools have their own unique demographics, funding, teacher-student ratio, etc. that define said school's identity. This research aims to analyze the aforementioned variables, as well as other variables listed on each school state report card, in order to create a model to predict any school's graduation rate.

In order to create this model, data was taken on all public high schools in Kentucky from the Kentucky Department of Education's School Report Card. Data were then narrowed down to only schools that had data available for all categories in the research. This left only 223 schools of the original 1477 to study. Regression analysis was then performed in Microsoft Excel and the statistical program R in order to make a model to predict the schools' graduation rates.

I dedicate this thesis to my parents, Roy and Angela Price, and my siblings, Alexander and Savannah Price, who have supported and helped me get to this point. I also dedicate this thesis to my professors who have given me the knowledge and encouragement to tackle this project.

ACKNOWLEDGEMENTS

I want to acknowledge Dr. Nicholas Fortune and all of the help he has given me along the way with this thesis. He has gone above and beyond to ensure that I am successful. I also want to acknowledge my friends Mary, Mattie, and Kate who have listened to endless drafts and presentations of this thesis along the way. I would also like to thank the Kentucky Department of Education for making their data available to students like me, allowing for projects like this to happen. Last but not least, I would like to thank the Mahurin Honors College at Western Kentucky University for their continued support throughout my college experience.

VITA

Education

Western Kentucky University - Bowling Green, KY May 2021
B.S. in Middle Grades and Secondary Math Education and Mathematics
Mahurin Honors College
Honors CE/T: *Regression Analysis: Graduation Rate in Kentucky Public High Schools*

Bullitt Central High School, Shepherdsville, KY May 2017

Professional Experience

Bowling Green Junior High/ Bowling Green High School January-May 2021
-Student Teacher

Kentucky Science Center – Louisville, KY June-August 2018/2019
-Counselor/ Teacher

Student Athlete Success Center – WKU Fall 2018 – Fall 2020
-Math Tutor

Governor's Scholars Program – Bellarmine University July 2020
-Resident Advisor

WKU Mathematics Department Ambassador Fall 2018 – Spring 2021

Bullitt County Public Schools/ Bowling Green Independent Schools Fall 2019 – Present
-Substitute Teacher

Awards/ Honors

Google Educator Level 1 Certified

President's Scholar/ Dean's Scholar

Brother of the Semester – Phi Sigma Pi, Beta Phi Chapter Fall 2018

Cherry Presidential Scholar Finalist

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The goals of education are to learn, grow, and ultimately start a career. An important part of that process is graduation – but what affects a student's ability to graduate? The combination of all students' success in graduating when looking at one particular school is defined as a graduation rate. Does a Variable that affects one school's graduation rate affect another's just the same? Kentucky's Public High School graduation rates vary across regions of the state (e.g., urban, suburban, and rural), provoking the question: can variables that make up a school's identity help predict what its graduation rate will be? The purpose of this study was to determine, via regression analysis, if different variables that make up a school's identity are able to significantly predict said school's graduation rate.

A school's graduation rate is important because it indicates not only what percentage of the students graduate in four years, but it also incites further exploration into a school's inner workings. Take for example if a school has a very high graduation rate; i.e., 95 percent or above. It would be beneficial to other schools with a lower graduation rate to understand exactly what the school with the higher rate is doing so that they can apply those initiatives to their own schools. In this study, different variables were analyzed to predict the graduation rates of schools. The variables that are able to accurately predict a school's graduation rate show us that this is something to investigate further in the school, as it has a significant effect on students' success in graduating.

The goal of this study was to analyze multiple variables in order to create an equation to predict any Kentucky high school's graduation rate based on said variables.

In addition to finding equations to predict graduation rates, this research aimed to find which of the combinations of variables has the best predictive ability for graduation rates. The null hypothesis of this study was that when all of the variables chosen are combined, there is no relationship with graduation rate to be a significant predictor. The research hypothesis of the study was that when all of the variables are taken into account, that combination will be the most accurate predictor of graduation rate, and that it will be a significant predictor. In this paper, there is a discussion on the variables that were chosen and why they were chosen over any others. Further, there is a discussion of the process of data organization, analysis, and a discussion of what the results of the study mean.

DATA COLLECTION AND ORGANIZATION

In order to begin this study, a reliable source of data for all schools in Kentucky needed to be found. Data for this study were taken from the Kentucky School Report Cards that are available on the Kentucky Department of Education (KDE)'s website [1]. These report cards were chosen because they are statistical documents that are required to be completed by every Kentucky school that receives state funding. This ensured that no public school would be overlooked in this study. The report cards contain information on a school's demographics, financials, rating in relation to other schools, etc. All of the data for any given Variable on a report card across all schools is available in the format of a downloadable spreadsheet.

Variables. On KDE's website there are 50 variables on the school report cards from which to choose to look at and analyze for each school. Appendix A shows all of these variables broken down into the categories in which they are listed on the website.

For the next step in this study, the variables to analyze needed to be determined. Analyzing all 50 variables would overcomplicate and slow down this study, therefore criteria were created to determine which of the variables would be used in the study.

The first criterion chosen was that the data provided must vary greatly from school to school in different regions of the state. For example, a school in urban Louisville may have a low student teacher ratio and a high non-white student population, while a school in rural eastern Kentucky may have exactly the opposite. To determine this variation across regions, a school from each of the following regions, urban Louisville, Eastern Kentucky, and suburban Bullitt County, were chosen as representatives of their areas to check if the data points for selected variables varied widely across these three regions. The variation for the variable chosen was checked by pulling the data for the three representative schools, and so long as each school varied from both of the two others by one standard deviation, then it was determined that the data was varied greatly enough from region to region.

The second criterion was that each needed to have data available for every public school. Some of the variables listed on KDE's website did not have data for certain schools and, therefore, could not be used in this study. The third criterion was that all of the variables chosen needed to have quantitative data, that is, numbers for their data rather than words, which is qualitative/categorical data. This is because on KDE's website, qualitative/categorical data were limited, and therefore there was more choice for variables when looking at quantitative data.

Lastly, the final criterion was that they represented different areas of a school's identity (i.e., no two variables would represent the same aspect of a school's identity).

For example, the number of gifted and talented students and the number of students in advanced coursework would not both be chosen because they represent a similar group of students. This selection of criteria is a small limitation to the study, which can be easily expanded for future endeavors. Future analysis can consider more variables from KDE's website with appropriate statistical analyses.

After all of these criteria were taken into account, five variables remained for the study. These variables were a school's student-teacher ratio (STR), the number of students on free and reduced lunch in a school (FRL), the total population of a school (TP), the total number of non-white students in a school (NW), and the spending per student of a school (SPS). All of these variables met the first two criteria above and were also determined by the researcher to be different enough from one another to analyze. The student-teacher ratio of a school is telling of the number of teachers in the school, the number of students on free and reduced lunch is telling of the economic status of the students in the school, the total population is telling of the size of the school, the number of non-white students is telling of the diversity of the school, and the spending per student is telling of the funding of the school.

The schools. After these five variables were chosen along with graduation rate, as this is the main focus of the study, the number of public schools was then examined. On the original data sheets taken from KDE's website, there were 1,477 schools listed. When the list of schools was examined, it was discovered that many did not have data listed for one or more of the five variables chosen or did not have a graduation rate listed (i.e., criterion 2 was not met). After schools were deleted for not having data available, there were 223 schools left to analyze. The primary reason that the list was reduced as much as

it was is because many of the schools deleted were not actual high schools. Many were technical schools, elementary/middle schools, juvenile detention centers, etc. These are places where students can attend school but cannot graduate.

Data organization. All of the variables chosen originally had data on individual spreadsheets downloaded from KDE's website. In order to simplify this analysis, all five of these sheets were compiled into one spreadsheet, which can be found in Appendix B. This spreadsheet was then used for all further analyses. In addition to this sheet containing the data for all six variables (the predictor/explanatory variables along with the response variable), additional sheets were created to hold each of the different variable combination regressions that would later occur. Each of these sheets were labeled with initials for their variables and a "vs. GR" for graduation rate. There are a total of 27 of these sheets for the 27 total combinations of the variables in groups of one, two, three, four, and all five variables against graduation rate. All of these sheets can be found in Appendix C.

DATA ANALYSIS

All data analyses for this study were conducted in Microsoft Excel.

Linear regression. The next step in the process of the data analysis for this study was to determine exactly what type of analysis needed to be run. The purpose of this study was to find a set of equations that could predict graduation rate based on given variables. For any equation, one needs a predictor/explanatory variable(s), coefficient(s) for these variable(s), and a response variable. The goal of this study was to find the graduation rate of a school, so this rate would be the response variable in the equations.

The predictor/explanatory variables would then be the variables mentioned earlier in this paper, such as student-teacher ratio and total population. Now that the variables were set, a mathematical process needed to be chosen to find the coefficients for these variables.

Simple linear regression is a mathematical model that takes a set of data, and through using a least squares regression, creates a linear model using the data set's X-variable to determine what the Y-variable is predicted to be [2]. This technique was beneficial for this study's analysis as that was exactly what the study aimed to do: create an equation that could predict graduation rate based upon given X-variables. Simple linear regression, however, only allows for one variable to predict the Y-variable. While this study did run each of the five variables individually against graduation rate, combinations of two, three, four and five variables also needed to analyze graduation rate, so a more advanced analysis technique than simple linear regression was needed.

Multiple linear regression. An expansion on simple linear regression called multiple-linear regression was then examined to determine if it could be used. Multiple linear regression expands on a simple linear regression in the sense that it allows the user to run an analysis with one or more variables at a time to determine the Y-value rather than limiting it to one variable as simple linear regression does. This is the mathematical process that was needed to run an analysis with the variables chosen for the study to result in a predicted graduation rate.

In its basic form, multiple-linear regression results in the following equation [3]:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon.$$

In this equation, $\hat{Y}$ is the resulting variable that the user wants to predict. In this study, the $\hat{Y}$-variable produced would be the predicted graduation rate based upon the variables chosen in that particular regression. In the equation, all of the β's represent different coefficients. The β_0 is the intercept of the equation, or in simpler terms, what the baseline graduation rate would be if all of the variables given in the rest of the equation had a value of 0. Baseline graduation rate means that each regression equation uses different variable combinations which result in unique equations. The rest of the terms in the regression equation are what change the value for an individual school based upon the data in the variables given.

The terms β_1, β_2, etc. represent the coefficients determined by the regression analysis that correspond to the given variables. The X's represent different variables used in the regression. Because this is a multiple-linear regression, there are multiple X's listed to represent the different variables chosen. The $\beta_n X_n$ variable listed at the end of the general form of the equation above represents the last Variable listed for a regression equation. Lastly, the ε represents any error in the equation. This estimated error is determined through analysis and given in the ANOVA table provided from the regression. For this study which is a multiple linear regression with five variables, this equation would have $\beta_0 - \beta_5$ and $X_1 - X_5$ where each subscript pair are adjacent to one another in the given equation. It would be of the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon.$$

This is the general form of the equation where all five variables are used. In the instance where all five are used, Table 1 provides what each element of the equation represents.

Table 1. *Five-Variable Explanatory Table.*

Y	Represents what the equation is equal to, in this study that is the school's graduation rate
β_0	The intercept, or "baseline" graduation rate
$\beta_1 - \beta_5$	Represent the coefficients taken from the regression table for the variables they are attached to
X_1	Student-teacher ratio for a school
X_2	Number of students in a school on free and reduced lunch
X_3	Total population of a school
X_4	Total non-white student population of a school
X_5	Spending-per-student of a school
ε	Error term

The descriptions given in the table are just for the given example using all five variables. In a two-variable regression equation, X_2 may not always represent the total number of students on free and reduced lunch, but rather the second variable being considered, so it is important to always look at what the equation represents. The variables will always be in order of what was mentioned first and second on the equation description.

In Microsoft Excel there is a tool called Data Analysis that was used for this study. This tool allows the user to run either a simple linear regression or a multiple-linear regression on data in a spreadsheet. Because the data for this study was already in the format of a spreadsheet, this tool was used to conduct all multiple-linear regressions for this study.

Part of the results from a regression are residuals. Residuals are the resulting difference from the predicted value given by the regression from the original data point [4]. Although residuals were not used in this study, they were calculated for every regression in this study in order to have the data available should any further analysis be conducted after this study on any individual school.

Once all the data are entered and the residuals are turned on, the "regression analysis" tool within Excel then outputs a list of tables. An example of the output table will be discussed in the results. The output tables contain a lot of information about the data and what the regression analysis found; however, there are a few things that were specifically focused on for this study. The main feature focused on from the resulting table after a regression was run is the column that is labeled "coefficients". An example of a "coefficients" table is shown in Figure 1.

	Coefficients
Intercept	93.80696945
X Variable 1	0.171114213
X Variable 2	-0.009518899
X Variable 3	0.004930639
X Variable 4	-0.006262223
X Variable 5	-6.84515E-05

Figure 1. Coefficients Output.

In the coefficients table shown, there are five X-variables. In this table, X-variables represent the same variables as these same variables do in Table 1. As mentioned, the coefficients are the B-values in the general form of the equation for a multiple-linear regression. This part of the resulting table is valuable because it provides the equation for this particular regression. Using the general form of the regression equation and the data provided above, we can now determine that the predictive equation using all of the variables in this study is approximately

$$\hat{Y} = 93.807 + 0.171X_1 - 0.010X_2 + 0.005X_3 - 0.006X_4 - (6.845 \times 10^{-5})X_5$$

While this is an example for just one regression, all resulting regression equations were determined in this manner. All regression equations for all combinations of variables can be found in Appendix D under "regression equation" for any combination.

Significance and predictability of the models. After the regression equation for all combinations of variables were found, it then had to be determined if these equations were significant in predicting the graduation rate of a school, or if they should not be used at all. In order to determine if an equation should be used or not, three elements given by the regression table were used: the *p*-values, the adjusted R^2 value, and the significance F statistic of a model (which was used to find the p-values).

A *p*-value is a statistical result that helps determine if a null hypothesis should be rejected or not based upon its statistical significance. According to simplypsychology.org, "The level of statistical significance is often expressed as a *p*-value between 0 and 1. The smaller the *p*-value, the stronger the evidence that you should

reject the null hypothesis" [5]. The cutoff for significance with a *p*-value is 0.05, meaning that if a *p*-value is above 0.05 that it is determined to not be statistically significant and therefore one should not reject the null hypothesis. For this study, if a *p*-value is above 0.05 this means that the combination of variables it is associated with does not significantly contribute to the regression equation. However, if a *p*-value is at or below 0.05, this means that it is statistically significant and one can reject the null hypothesis. This means that in the combination of variables, at least one of the said variables contributes significantly to the prediction of the response variable. In all regressions run for all variables, each variable has its own individual *p*-value. This is notable because it is possible that within the same regression, one variable may be significant while another is not.

Another important element given by the regression table for analysis is the R^2 value. According to people.duke.edu, "R^2 is the 'percent of variance explained' by the model. That is, R^2 is the fraction by which the variance of the errors is less than the variance of the dependent variable" [6]. An R^2 value is a number between 0-1 that ideally would be as close to 1 as possible to show the least variance in error in the data set. However, for this study, an adjusted R^2 value was used. This is because an adjusted R^2 value accounts for not only the variance of the data, but also the number of variables used.

After the *p*-values and R^2 value of a regression are analyzed, a deeper look into the regression equation can occur. As previously mentioned, if a variable has a low *p*-value (below 0.05), this means that it is statistically significant to the model. Also mentioned is the fact that a low R^2 value (lower than 0.5) means that the data has high

variance in relation to the given equation. If the provided significance p-value from the regression table for any given equation is below 0.05, then it was determined that the data is significant and therefore, the regression equation is significant. This however does not mean that the low R^2 value should be ignored; but rather that the equation calculated is the best predictor available for the given data.

In addition to this, it is worth noting that coefficients can have positive or negative values. In this study, assuming all other variables stayed constant, a negative coefficient means that the variable lowered the graduation rate of a school in that specific model. Similarly, a positive coefficient was interpreted as the variable raising the graduation rate of a school. For example, if the total population of a school and the spending per student of the school were used run in a regression together, and the total population had a positive coefficient, but the spending per student had a negative coefficient, this would be interpreted to mean that in this particular regression, the higher the total population of a school, the more the graduation rate raises while the greater the spending per student, the lower the graduation rate, when the other variables is held constant. In the next section, the results from the analyses are discussed.

RESULTS

In this study, a full regression analysis was run for each combination of variables which resulted in an equation, p-value and a R^2 value. These results can be found in Appendix D. The purpose of this study was to find a model via regression analysis that is a significant predictor of a school's graduation rate based on the variables and data

provided. Based on these criteria to have a significant regression, the significant combinations of variables are all listed under Appendix D indicated by a highlighted value. Of all 27 regressions run, 15 meet the criteria to be determined as significant.

Recall, the null hypothesis of this study was that when all of the variables chosen for the study are combined into one regression, it will be determined that the combination of variables in the regression equation is not a significant predictor of graduation rate. The alternative hypothesis of this study was that when all of the variables are combined into one regression, it will be a significant predictor of graduation rate. Below in Figure 2 is the regression output when all variables were used.

Regression Statistics	
Multiple R	0.532328609
R Square	0.283373748
Adjusted R Square	0.266785177
Standard Error	3.912163847
Observations	222

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	1307.238175	261.4476351	17.08246923	3.15623E-14
Residual	216	3305.885608	15.30502596		
Total	221	4613.123784			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	93.80696945	2.371744961	39.55187888	6.7362E-101	89.13224241	98.4816965	89.13224241	98.48169649
X Variable 1	0.171114213	0.123258535	1.388254482	0.166489666	-0.07182928	0.41405771	-0.07182928	0.414057706
X Variable 2	-0.009518899	0.001967439	-4.838218308	2.49051E-06	-0.013396735	-0.00564106	-0.01339674	-0.005641062
X Variable 3	0.004930639	0.001095498	4.500818235	1.10628E-05	0.002771404	0.00708988	0.002771404	0.007089875
X Variable 4	-0.006262223	0.001558061	-4.019240041	8.06525E-05	-0.009333174	-0.00319127	-0.00933317	-0.003191272
X Variable 5	-6.84515E-05	6.64574E-05	-1.030006056	0.304159126	-0.000199439	6.2536E-05	-0.00019944	6.25365E-05

Figure 2. Regression Output for all Variables.

As seen in Figure 2, when all five variables are used for the regression, the adjusted R^2 value is 0.267, a very low result. The *F*-statistic indicates that the equation is significant. When we look at *p*-values, *X*-variables 2, 3 and 4 are significant, however, 1 and 5 are not. Therefore, while the regression equation for all five variables is significant, some of the variables are not. Because of this, we cannot reject the null hypothesis

because certain other regression equations are significant and all variables combined is not. But this leads to further analysis: which combination of variables is the best predictor of graduation rate?

Appendix D contains the list of all the regression combinations which were determined to be significant. Because the regression which is the best predictor of graduation rate must have a significant equation, only combinations from this list were considered. Because *p*-value was previously analyzed in order to be on this list, the adjusted R^2 value was the deciding variable for which regression was the best predictor. Ideally, for a regression equation to be a good predictor of graduation rate, the adjusted R^2 value should be as close to 1 as possible, however, in this study none of the regressions had an adjusted R^2 value above 0.3. This is not a good adjusted R^2 value, meaning that this study cannot conclude that any of the predicted values will be exact or precise predictions. Despite this fact, the variable combination with the highest adjusted R^2 value was determined, and its regression output is shown below.

Regression Statistics								
Multiple R	0.520874824							
R Square	0.271310582							
Adjusted R Square	0.261282746							
Standard Error	3.926815879							
Observations	222							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	3	1251.5893	417.1964	27.05574581	6.48229E-15			
Residual	218	3361.534483	15.41988					
Total	221	4613.123784						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	95.24488216	0.640728114	148.651	4.2536E-221	93.98206753	96.5076968	93.98206753	96.50769679
X Variable 1	-0.009984592	0.001954742	-5.10788	7.10064E-07	-0.013837205	-0.00613198	-0.013837205	-0.006131979
X Variable 2	0.005868729	0.000973176	6.030494	6.91875E-09	0.003950692	0.00778677	0.003950692	0.007786766
X Variable 3	-0.007145599	0.001492999	-4.78607	3.13567E-06	-0.010088159	-0.00420304	-0.010088159	-0.004203038

Figure 3. Results of FRL and TP and NW vs. GR Regression

As seen in Figure 3, it was determined that the regression with the highest adjusted R^2 value while also having all significant p-values under 0.05 was the combination of the number of students on free and reduced lunch, the total population of a school, and the number of non-white students. In this regression table, X-variable 1 represents the number of students on free and reduced lunch, X-variable 2 represents the total population of a school, and X-variable 3 represents the number of non-white students. The regression table also provides coefficients for these variables, allowing us to create an equation for this predictive model. The equation is

$$\hat{Y} = 95.245 - 0.010X_1 + 0.006X_2 - 0.007X_3.$$

This equation can be interpreted as follows. The number 95.245 represents the baseline graduation rate. This means that when these three variables are considered, the estimated graduation rate when $X_1 = X_2 = X_3 = 0$ is about 95.245%. Both the variables representing the number of students on free and reduced lunch and the number of non-white students have negative coefficients. This can be interpreted as when FRL and NW are taken into account in the regression, overall, the higher the rates for these variables in a school, the lower the graduation rate will be when all other variables are held constant. Alternatively, the variable representing total population has a positive coefficient. This means that as the TP of a school gets higher, the graduation rate will rise assuming that the free and reduced lunch variable and the non-white student variable are held constant.

As seen in Figure 3, the F-statistic and p-values for all of the variables in this regression combination are well below the 0.05 cutoff for significance, therefore, it can

be determined that this regression equation is significant. The next value in the regression table to look at is the adjusted R^2 value. Ideally, the adjusted R^2 value for a regression equation would be as close to 1 as possible, however, as mentioned earlier in the analysis of this study, none of the adjusted R^2 values for the regressions in this study were above 0.3. The value of 0.261 for this combination as seen in Figure 3 was the highest of any regressions with a significant equation, therefore, making FRL and TP and NW vs. GR the best predicative equation given by this study. This combination of variables as the most significant predictor of graduation rate makes sense when the combination of all variables is analyzed. The reason that the combination of all variables is not on the list of significant regressions is because the *p*-values for student teacher ratio and spending per student are not below the 0.05 cutoff for significance and therefore, because not every variable in the equation was significant it was not included. When you remove these two variables from the combination, you are left with the three listed above, which is the reason that this particular combination appears to hold true to being the most significant predictor of graduation rate.

As mentioned earlier, although this study has a combination of variables that is more statistically significant than the rest, in the grand scheme of the study, it appeared that none of the variable combinations are significant predictors of graduation rate although they are still statistically significant. In order to test this assumption, a test for practical significance was conducted. According to statisticshowto.com, "Practical significance relates to whether a result from a statistical hypothesis test is useful in real life. It is a way to address some of the limitations with traditional testing and answers the question: Do your results have real life applications and meaning?"[8]. Overall, practical

significance allows the researcher to determine if their results are a good predictor for the application of their results, in this case, if this regression combination is a practical predictor of graduation rate. In order to determine if the most significant variable combination from this study has practical significance, an effect size test was conducted.

An effect size test determines if the combination of variables has practical significance. Note that in order for an effect size test to be conducted, the regression must first have statistical significance determined by the *p*-value of the variables [8]. The first component of an effect size test is determining the pooled standard deviation, or the standard deviation across all variables in the regression. The formula for this is

$$S_p = \sqrt{\frac{(n_1 - 1)(s_1{}^2) + (n_2 - 1)(s_2{}^2) + \cdots + (n_m - 1)(s_m{}^2)}{n_1 + n_2 + \cdots + n_m - m}}.$$

In the formula above, S_p is the pooled standard deviation that we are looking for, $n_1 \ldots n_m$ represent the number of data points in each variable in the regressions' set, $s_1 \ldots s_m$ represent the sample standard deviations of each of the variable's data sets, and m represents the number of variables considered in regression. When the data for the variable combination FRL and TP and NW vs. GR was placed into this equation, the results were as follows

$$S_p = \sqrt{\frac{(223 - 1)(249.125^2) + (223 - 1)(476.516^2) + (223 - 1)(255.492^2)}{223 + 223 + 223 - 3}}$$

$$= 208.048.$$

After the pooled standard deviation is found, the next step in determining the effect size is to find Cohen's *d*-value. This is the value that will later help determine if the data is practically significant or not. The formula to find this value is

$$d = \frac{\bar{X}_1 - \bar{X}_2 - \cdots - \bar{X}_n}{s_p}.$$

In this formula, $\bar{X}_1 - \bar{X}_n$ are the sample means of the variables in the regression combination from largest to smallest. The resulting *d*-value for the regression combination FRL and TP and NW vs. GR is

$$d = \frac{867 - 478 - 187}{208.048} = 0.971.$$

Now that the *d*-value has been obtained, it can be analyzed by comparing it to the values in Table 2.

Table 2. *Cohen's d-value Interpretation* [2].

Cohen's d-value	**Interpretation**
0-0.2	Little to no effect
0.21-0.5	Small effect size
0.51-0.8	Medium effect size
0.81 or more	Large effect size

As seen above, the *d*-value for the regression combination was 0.971. This means the *d*-value in the large effect size category, meaning that not only is it statistically significant, but also practically significant. This means that this regression combination can be used to predict a school's graduation rate.

In addition to this, in order for a combination of variables to be a good predictor of graduation rate, the *p*-value should be lower than the 0.05 cutoff for significance as well as have an adjusted R^2 value above 0.5 (ideally as close to 1 as possible). However, none of the regressions in this study meet this criteria having both a low *p*-value and a high adjusted R^2 value. Because equations were determined to be significant through the *p*-value and have practical significance, the adjusted R^2 value indicates that the equation for the regression is a good predictor of graduation rate, but due to the high variance in the data given for the study, the data points do not relate well with these significant equations.

Although the Variable combinations in this study were not great predictors of graduation rate, the data still revealed important information. This study revealed that spending per student and student teacher ratio are not necessarily relevant in determining a schools graduation rate, while the total population, number of non-white students, and number of students on free and reduced lunch can be. If this study were to ever be expanded, of the five variables chosen, the three that were determined to be more significant should be used moving forward.

CONCLUSION

Before this study was conducted, it was predicted that the combination of all five variables chosen: student teacher ratio, the number of students on free and reduced lunch, total population, number of non-white students, and spending per student, would be the best predictors of graduation rate for any given Kentucky public high school. However, through the process of this study, it was found that none of the combinations of variables were good predictors of graduation rate; however, of the variables, the combination of free and reduced lunch, total population, and the number of non-white students was the most significant predictor. This was based on the fact that the p-values for this combination were all significant, and this combination held the highest adjusted R^2 value of any combination of variables.

As seen in the results section analysis of this equation, the regression equation for this Variable combination has two negative coefficients and one positive. Because the coefficients for the variables of the number of students on free and reduced lunch and the number of non-white students were negative, it was determined that based on this regression, these two variables have a negative relationship with a schools' graduation rate, i.e., the more students on free and reduced lunch in a school and the more non-white students, the lower the overall graduation rate of the school. Conversely, the variable of total population had a positive coefficient, meaning that it has a positive relationship with a schools' graduation rate, i.e., the larger the total population of a school the higher the graduation rate overall. This analysis is very telling because should any further analyses be conducted on variables effect on graduation rate.

Although it was found that none of the combinations of variables were good predictors of graduation rate, the information from this study is still valuable. If this study were to ever be expanded upon, the three variables given as the most significant predictors could be used in further analysis while the other two could likely be excluded from further analysis. As mentioned before in this paper, the variables chosen for this study were only five of 50 total variables available on KDE's website. If this study were to be expanded upon in the future, the three variables of the number of students on free and reduced lunch, the total population, and the number of non-white students should be included in further analysis along with other new variables which were not included in this study. This study could also be expanded beyond just five variables to include more combinations and regressions that may yield more detailed results as the combinations of variables become larger. There are many different ways that this study could be expanded upon using the same process described in this analysis in order to find a combination of variables that yields the most significant predictor of graduation rate for public high schools in the state of Kentucky.

REFERENCES

[1]: Kentucky Department of Education. (2019). School Report Card Data Sets. Retrieved from Kentucky Department of Education: https://openhouse.education.ky.gov/Home/SRCData

[2]: Penn State Statistics Department. (2020). STAT 462. Retrieved from Penn State: https://online.stat.psu.edu/stat462/node/91/

[3]: ReliaWiki. (2020). Multiple Linear Regression Analysis. Retrieved from ReliaWiki.com: http://reliawiki.org/index.php/Multiple_Linear_Regression_Analysis

[4]: StatisticsHowTo.com. (2020). Residual Values. Retrieved from statisticshowto.com: https://www.statisticshowto.com/residual/

[5]: SimplyPsychology.com. (2020). What a p-value tells you about statistical significance. Retrieved from SimplyPsychology.com: https://www.simplypsychology.org/p-value.html

[6]: Duke University. (2018). What's a good value for R squared? Retrieved from people.duke.edu: https://people.duke.edu/~rnau/rsquared.htm

[7]: StatisticsHowTo.com. (2019). F statistic/ F Value: Simple Definition and

Interpretation. Retrieved from statisticshowto.com: https://www.statisticshowto.com/probability-and-statistics/f-statistic-value-test/#:~:text=If%20you%20get%20a%20large,of%20all%20the%20variables%20together.

[8]: StatisticsHowTo.com. (2020). Practical Significance. Retrieved from statisticshowto.com: https://www.statisticshowto.com/practical-significance/

APPENDICES

Appendix A

https://docs.google.com/document/d/1UI_HBX-xwn_PW5vXk83dWdj7t5OIKkQvD8hdo4szMOM/edit?usp=sharing

Appendix B

https://docs.google.com/document/d/1ICIrrM0tGyMv2bYBY86xM8pg6ExZyfDGPHKL1r_QYMw/edit?usp=sharing

Appendix C

https://drive.google.com/file/d/1go_DuSO3T-J031lY2l_w_3t48FS3rOJc/view?usp=sharing

Appendix D

https://docs.google.com/document/d/1qUmO-QlIvzzBLG-eqiZyyRRpdhFv0YXsRj7E5E7WIz8/edit?usp=sharing