Predicting Average 8th Grade Test Scores:

An analysis of Georgia school districts in 1996

Shreya Ganeshan
STAT 4360
Final Project
6 December 2016

I. SUMMARY

The goal of this analysis is to effectively and accurately predict average 8th grade test scores for a school district in Georgia using ten different variables. This list of predictors consists of both numerical and categorical variables, which impact the interpretability of the results obtained through this empirical work.

This paper first explores summary statistics of all variables in the dataset, “gasd96.txt.” Numerical variables are displayed in a tabular format, while categorical variables are resented in frequency tables and bar plots. Visual representations of the relationships between the response variable of 8th grade test scores and each individual predictor variable demonstrate the need to manipulate numerical and categorical predictors differently, using scatterplots and boxplots, respectively.

Before proceeding with regression analysis in order to predict 8th grade math and reading test scores, all necessary assumptions were checked using statistical methods. When checked for normality, the response variable seems to present a fairly normal distribution, eliminating the need for any log transformation of this variable. A backward procedure for variable selection based on p-value significance exposes the most useful parameters for predicting 8th grade test scores – including, average 3rd test scores, average 5th grade test scores, school district enrollment, population density, metropolitan area indicator, medium-enrollment district indicator, and the proportion of students eligible for free lunch. Rather than including all variables given in the dataset to predict the response variable, this final model eliminates the effects of any over-prediction as found in the null (or full) model.

A thorough analysis of the dataset reveals that only seven of the ten provided variables significantly contribute to the prediction of average 8th grade test scores in 1996. Challenges in interpretation arise from the inherent correlation between certain variables – namely, population density, metropolitan indicator, medium-enrollment district indicator, and large-enrollment indicator – which all describe the size and location of a school district. Since, the dataset does not differentiate test scores by subject, interpretation of the model must forego a degree of specificity with regard to which subject can better be explained. Given the diversity of variable and correlations, this analysis attempts to elicit a realistic and relatable interpretation of how different parameters can be employed to predict 8th grade student outcomes.

II. INTRODUCTION

The goal of this analysis is to identify whether specific variables related to prior student achievement and school district size, location, and income-levels are useful predictors of average test scores among 8th grade students in the state of Georgia in 1996. As an explanatory analysis, the regression model investigated in this paper can be utilized to measure the strength of relationships between independent and dependent variables rather than assessing causality. The magnitude of correlation suggested by the model could inform subsequent opportunities to explore cause/effect relationships.

In the "gasd96.txt" dataset, there are 174 observations, corresponding to the total number of districts (159 counties and 15 independent systems) in Georgia in 1996. The variables explored in this analysis can be found in the table below. Numerical variables include numeric responses in the dataset. Among numerical variables, discrete variables can only be interpreted in whole units, while continuous parameters can be interpreted in fractions or decimals of units. Categorical variables are represented as binaries coded as 1 if the observation satisfies the categorical parameter and 0 if the observation does not. For example, a district located within a metropolitan area would be denoted as: IMTR = 1.

Table 1: Variables

Abbreviation	Description	Classification
COUNTY	School district name	Categorical
T8	Average 8th grade test Score (math & reading)	Numerical (continuous)
T3	Average 3rd grade test score (math & reading)	Numerical (continuous)
T5	Average 5th Grade Test Score (math & reading)	Numerical (continuous)
ENRL	Total number of students enrolled in school district	Numerical (discrete)
DENS	Population density in school district	Numerical (discrete)
IMTR	Indicator of metropolitan area	Categorical (1 = Yes, 0 = No)
IM	Indicator of medium-enrollment district	Categorical (1 = Yes, 0 = No)
IL	Indicator of large-enrollment district	Categorical (1 = Yes, 0 = No)
PLUN	Proportion of students in district eligible for free lunch	Numerical

After assessing and checking relevant regression assumption, this paper proceeds with a multiple linear regression model of first degree, non-interacting variables. The parameters included in the final regression follow a backward variable selection process based on statistical significance of p-values. Then, the paper further vets this final model through two processes of identifying outliers and assessing the need for transformations of the regression parameters. At last, the model is checked for time dependence and heteroskedasticity.

III. DATA SUMMARY

Table 2: Summary Statistics of Numerical Variables

Variable Names	Mean	Median	STDV	Min	Max	Q1	Q3	IQR
T8	95.207	95.5	18.158	45	138	84	107.5	23.5
T3	103.052	104	20.377	52	157	89	116	27
T5	100.207	100	17.885	44	140	88.25	114	25.75
ENRL	7726.586	3437	13836.505	457	90311	2006	7159.5	5153.5
DENS	181.59	62.915	356.119	5.75	2198.27	33.44	148.335	114.895
PLUN	0.497	0.504	0.166	0.065	0.876	0.399	0.614	0.215

Table 3: Frequencies of Categorical Variables

IMTR	Frequency	IM	Frequency	IL	Frequency
Metro Area	50	Medium Enrollment	67	Large Enrollment	19
Non-Metro Area	124	Non-Medium Enrollment	107	Non-Large Enrollment	155

Table 2 presents a summary of each numerical variable to offer preliminary insight into the nature of the observations captured by each factor. As measures of center, the mean (average of all observations given a variable) and median (the middle observation given an ordered set of observations for each variable) can be compared to determine the overall shape of each variable distribution. The standard deviation (distance from the mean), minimum, and maximum values provide insight into the spread or dispersion of each variable distribution. The first quartile, third quartile, and interquartile range provide information on the extreme values of each variable.

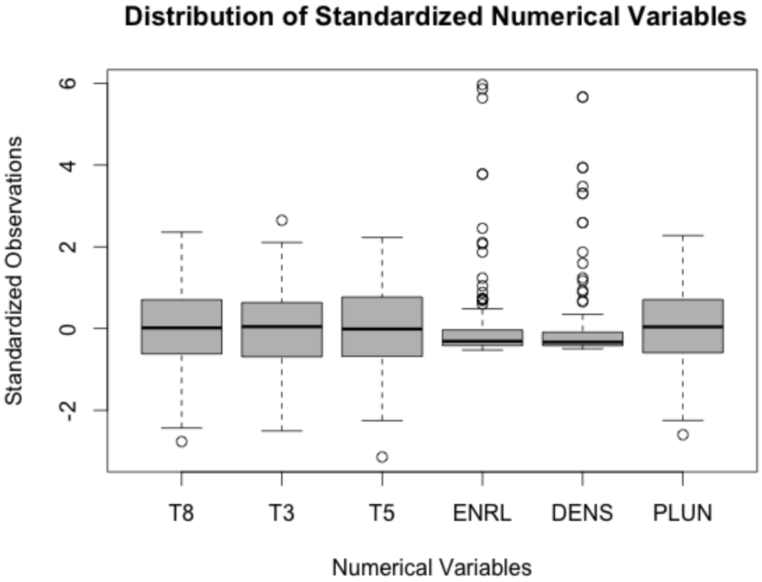
From Table 2, it can be observed that the median and mean of T8 are relatively close, indicating a relatively symmetric (perhaps normal) distribution. The mean of T3 is slightly less than its median, indicating possible left skewness of T3's distribution. The mean and median of T5 are quite close, also indicating a relatively symmetric distribution. The mean of ENRL is much larger than its median, pointing to significant right skewness, also evident from ENRL's large standard deviation. The same skweness applies to the distribution of DENS. However, PLUN's mean and median are relatively close, implying a fairly symmetric distribution.

Table 3 displays the frequencies of all binary categorical variables, since it does not make sense to calculate traditional summary statistics for these parameters. From the variable IMTR, it can be observed 74 more school districts in Georgia in 1996 were located in non-metropolitan areas. There were 40 more school districts classified as non-medium enrollment than medium-enrollment districts. Likewise, there were 136 more non-large enrollment districts than large-enrollment districts.

Similar to Table 2, Figure 1 individually summarizes individually quantitative variable distributions, but visually. Because all variables in the graph are all standardized, it is only practical to compare the spreads of variable observations, rather specific observations. The process of standardizing involves calculating individual observations' distances from their respective variables' means and dividing that variable's standard deviation. This allows for a unit-less comparison of spread. With regard to the boxplot features, the circles above the box tails represent outliers or data values unusually distant from the mean, based on the first (Q1) and third (Q3) quartiles as well as the interquartile range or IQR (difference between Q3 and Q1) represented in Table 2. The threshold for outliers can be determined by multiplying the IQR by 1.5; any observations above or below this (1.5*IQR) distance from the mean is considered an outlier. Outliers can cause concern in model analyses because they are highly unusual observations and can distort results. On the other hand, outliers an also offer helpful insight if they corroborate the conclusions drawn based on other variables or provide avenues for further directions the study can take.

Given the significant number of outliers contained in DENS and EMRL, it becomes clear that their distributions are highly skewed right. Transforming these two variables with logarithms or exponents can help correct the skeweness of the variable distributions and presence of outliers, providing room for more accurately assessing relationships and correlations in the model.

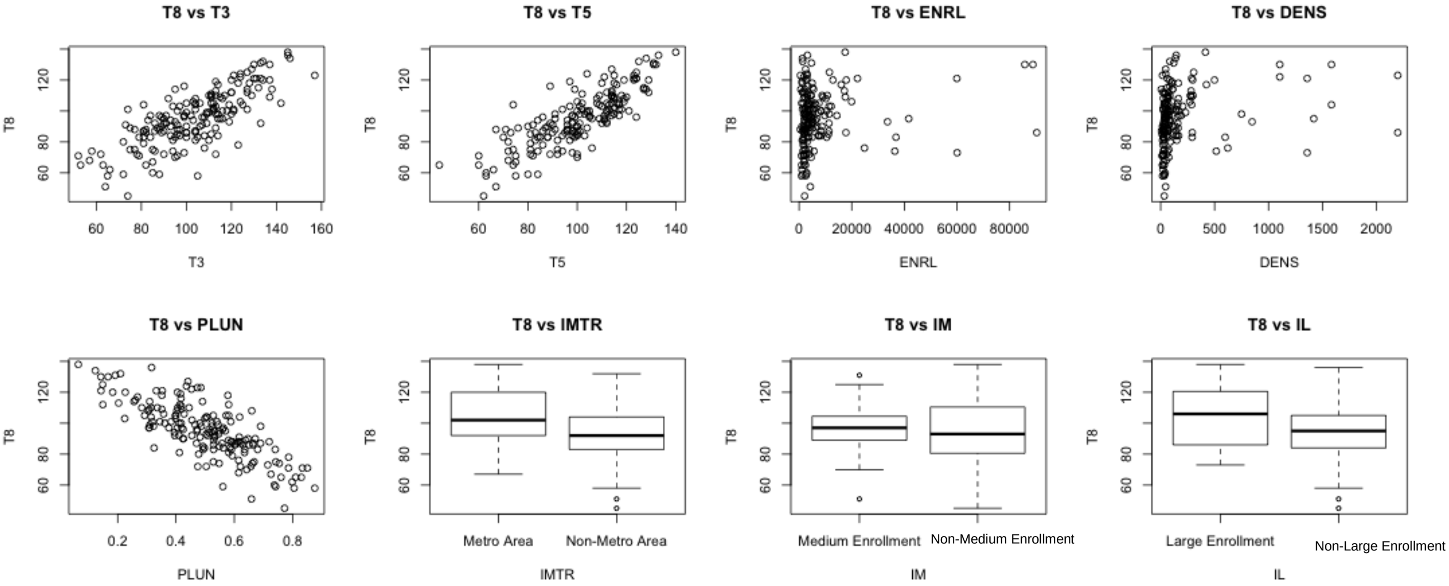
Figure 1: Standardized Box Pot of all Variables



On the next page, Figure 2 displays relations between T8 and all predictor variables, both quantitative and categorical. Between T8 and T5 as well as between T8 and T3 there are positive relatively positive linear relationships, meaning that as 3rd grade and 5th grade scores increase, at a relatively constant rate. Or, high values of one variable correspond with high values of the other variable. There is an obvious negative linear association between T8 and PLUN, indicating as the proportion of students in a district who are on free (typically from lower income backgrounds) increases, average 8th grade scores decreases. However, the relationships between T8 and DENS and T8 and ENRL initially seemed to display no obvious correlation due to their distinctly skewed right distributions (evident in Figure 2). In Appendix III, there are plots of T8 vs ln(ENRL) as well as T8 vs ln(DENS), whose positive linear relationships confirm that these two variables could be useful in helping explain average 8th grade test scores.

Unlike the numerical variables, the categorical variables IMTR, IM, and IL require boxplots rather than scatterplots. The binary responses values of “0” and “1” were recoded to represent the booleans relevant to each variable. For example, a value of “IMTR = 0” was recoded as “IMTR = ‘Non-Metro Area.’” From the boxplots, it can be observed that IMTR, IM, and IL have potential to be useful predictors of T8 because the variables seem to have fairly different means and medians. The differences in ranges and standard deviations of IMTR, IM, and IL do not impact whether or not the variables will help make accurate predictions of 8th grade scores.

Figure 2: Plots of T8 versus Each Variable



IV: ANALYSIS

In this section, a linear regression model will be explored used to predict average test scores based on a host of explanatory parameters. Regression analyses measure the relationship between variables and help quantify their predictive ability. After a statistical vetting process, this paper will recommend a regression in which all of the variables are useful in predicting average 8th grade test scores. Each predictor variable will bear a coefficient whose direction and magnitude will represent that variable’s respective relationship with T8, ceteris perebis (all else equal). Regressions ultimately provide equations to plug in variable values to yield a mean T8 value.

From regression outputs, we can also determine what percent of the variation in a response variable can be attributed to the explanatory variables in the model. The goal of this section is to make such claims about predicting average 8th grade scores.

The most effective and efficient and regression models must satisfy a series of assumptions about the data of interest. Only if the four assumptions included below are satisfied can a linear relationship be drawn based on the "gasd96.txt" dataset. We will explore these assumptions throughout the course of this analysis section.

- Normality: The response variable must be fairly normally distributed.
- Homoskedasticity or constant variance: All random variables in the model must have the same finite variance, or residuals vary evenly around all predicted values.
- Random and Independent Assignment: All observations in the model must comprise a random sample and were taken independently.
- No multicolinearity: The variables must be independent from each other.

Table 4 provides correlation matrix, which includes the correlation coefficients of all pairs of numerical variables in the data. The closer a coefficient is to 1 or -1, the stronger the linear association. Ideally, pairs of T8 and predictor variables would have strong linear correlations, but any strong linear correlation among explanatory variables should raise a red flag. In cases in which pairs of two independent variables have high correlation coefficients, only one of the two variables should be included in the regression equation because there is evidence that they both provide very similar information in predicting the response variable. Thus, including the most statistically significant (based on low associated p-values) in the model make sense, but including both will provide redundant information and distort the model's predictive ability. The diagonal of the table, representing a variable's perfect linear relationship with itself, will always be equal to 1. Also, it is worth noting that correlation coefficients can only be calculated among like variable types (i.e., numerical on numerical or categorical on categorical, not numerical on categorical). Table 5 provides correlation coefficients between categorical variables.

From Table 4, the only causes for concern seem are: the relationship between T3 and T5 and between ENRL and DENS. These variables are very strongly, positively correlated – with correlation coefficient values of 0.84 and 0.75, respectively. Though it is possible that T3 and T5 or both ENRL and DENS could remain in the final model, there is demonstrated evidence that the inclusion of both variables could cloud the regression's accuracy.

Table 4: Correlation Matrix of All Quantitative Variable Pairs

	T8	T3	T5	ENRL	DENS	PLUN
T8	1	0.77	0.82	0.16	0.23	-0.78
T3	0.77	1	0.84	0.18	0.23	-0.68
T5	0.82	0.84	1	0.24	0.24	-0.79
ENRL	0.16	0.18	0.24	1	0.75	-0.24
DENS	0.23	0.23	0.24	0.75	1	-0.19
PLUN	-0.78	-0.68	-0.79	-0.24	-0.19	1

Table 5: Correlation Matrix of All Categorical Variable Pairs

	IMTR	IM	IL
IMTR	1	0.1	0.51
IM	0.1	1	-0.28
IL	0.51	-0.28	1

The first and foremost assumption of the regression model, normality, demands an analysis of the distribution of the response variable and the distribution of residuals or error values. The first method, distribution of the response variable, can be observed without constructing a linear regression model. Figure 3 displays a relatively uniform density distribution, and Figure 4 displays a frequency distribution that mirrors a normal bell curve.

Figure 3: Density Distribution of T8

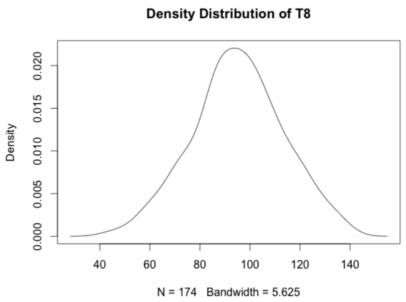
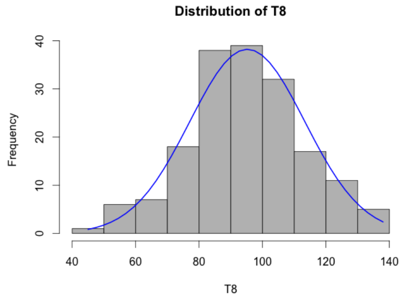


Figure 4: Frequency and Normal Distribution of T8



Though the above two graphs demonstrate that the response variable T8 is normally distributed, the normality is not officially satisfied until a Shapiro-Wilk Test is performed. This procedure tests the null hypothesis that the T8 sample is taken from a normally distributed population. The resulting p-value of 0.8412 means that this null hypothesis cannot be rejected. Thus, we assume that T8 is normally distributed and we can continue on with the rest of the analysis.

With the central normality assumption satisfied, we can proceed with the development of a regression model to predict average 8th grade test scores. A full model containing all numerical and binary categorical variables is included below, where β_i from *i* =1,..., 8 are the parameter estimates.

$$T8_i = \beta_0 + \beta_1 T3_i + \beta_2 T5_i + \beta_3 (ENRL)_i + \beta_4 (DENS)_i + \beta_5 IMTR_i + \beta_6 IM_i + \beta_7 IL_i + \beta_8 PLUN_i + \varepsilon_i$$

Preliminary regression outputs reveal that not all of these variables are statistically significant or useful as predictors. In order to determine which parameters to include in the regression, this paper utilizes an exhaustive selection procedure based on backward selection. This procedure removes variables from the regression in a step-by-step process beginning at the end of the full model. At the conclusion of this vetting process, the parameter variables which remain are: T3, T5, ENRL, DENS, IMTR, IM, and PLUN. The resulting final model is:

$$T8_i = \beta_0 + \beta_1 T3_i + \beta_2 T5_i + \beta_3 ENRL_i + \beta_4 DENS_i + \beta_5 IMTR_i + \beta_6 IM_i + \beta_7 PLUN_i + \varepsilon_i$$

An ANOVA F-test, in which all the four parameters have statistically significant F-values, confirms that the predictors in the model are statistically significant, with p-values less than a 0.01 level of significance. Table 6 displays this significance. Thus, the seven variables – T3, T5, ENRL, DENS, IMTR, IM and PLUN – are valid and useful predictors of average 8th grade test scores.

Table 6: Results from Final Model ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
T3	1	33563.0890	33563.0890	404.7567	0.0000
T5	1	5856.9066	5856.9066	70.6318	0.0000
ENRL	1	38.9054	38.9054	0.4692	0.4943
DENS	1	218.1750	218.1750	2.6311	0.1067
IMTR = Metro-Area	1	3.1849	3.1849	0.0384	0.8449
IM = Medium-Enrollment	1	369.0808	369.0808	4.4510	0.0364
PLUN	1	3224.2192	3224.2192	38.8827	0.0000
Residuals	166	13764.9909	82.9216	NA	NA

Now, that a regression model has been developed, we must check the remaining regression assumptions. From the "Normal Q-Q" plot in Figure 5, residual diagnostic plots demonstrate that the standardized residuals display normality. According to the "Residuals vs. Fitted" plot, the residuals or model errors are fairly evenly scattered around a mean of zero, demonstrating constant variance. There is no particular clustering or funneling effect present among residuals in this plot; therefore, the independence assumption is also satisfied. In the "Residuals vs. Leverage" plot, there are no significant outliers or influential points that would dramatically distort the data. In fact, the highest Cook's Distance in the model is 0.08, which is significantly lower than the threshold for concern (usually 1). Graphic confirmation that a linear regression model is can be seen in "Observed T8 versus Predicted T8" graph in Figure 7.

These claims can also be quantified using a Durbin Watson Test and Breusch-Pagan Test. The Durbin-Watson procedure checks the presence of autocorrelation in a regression, which causes the residuals in the current period (1996) to be correlated with residuals in other years (both previous and subsequent). The test statistics is 2.382, which is fairly close to the "no autocorrelation" threshold of 2. The Breusch-Pagan Test checks the residuals for non-constant variance, and according to a Chi-Square estimate of 2.465 and p-value of 0.1164, the residuals do not display heteroskedasticity.

Figure 6: Residual Diagnostic Plots

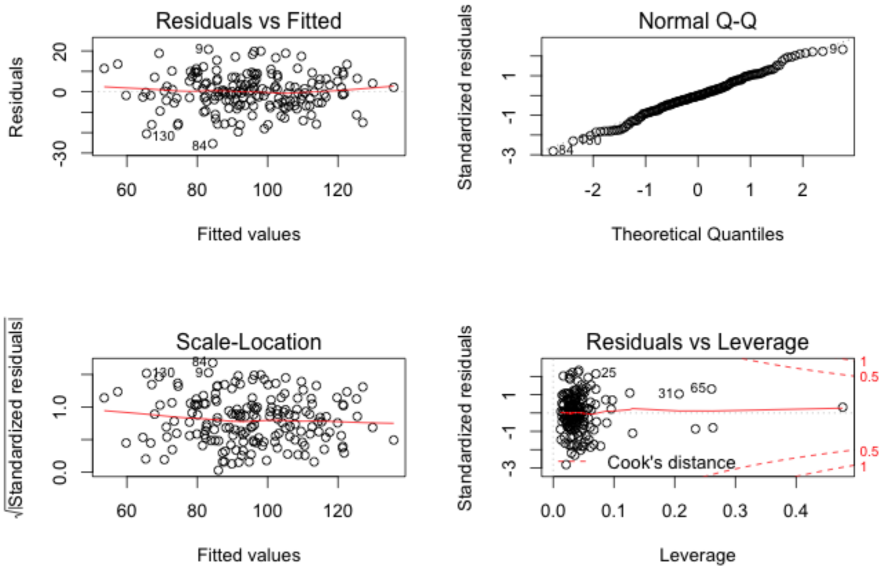


Figure 7: Observed versus Predicted Values

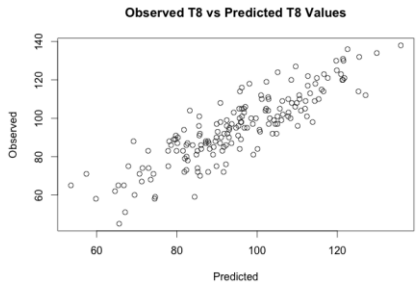


Table 7: Model Comparison

Model	Res.Df	Adjusted R^2
Null	173	0
Full	165	0.7479
Final	166	0.7485

We must check one final assumption – multicolinearity. We can see if the explanatory variables are independent from one another using variance inflation factors. None of the explanatory variables have variance inflation factors above 10 (the typical threshold for multicolinearity), thus the parameters are all independent from one another.

Table 7 above provides a simple tool to compare three different regressions. The “null” model represents a regression of T8 on 1, yielding the expected value or mean of T8. The “full” model contains all quantitative and numerical variables. And the “final” model contains: T3, T5, ENRL, DENS, IMTR, IM, and PLUN. The Adjusted R^2 value is the percent of variation in T8 that can be attributed to the model’s explanatory variables, after correcting for inflation due to adding multiple parameters to the model. Though the full model has a higher Adjusted R^2 value compared to the final model, the numbers are so close that we choose the model that it is evident that the seven variables in the final do a sufficient job at explaining the variation in T8. Therefore, for simplicity and interpretability purposes, we can settle with the final model.

The final regression model can be expressed in the following equation:

$$\widehat{T8} = 64.13 + 0.1810(T3) + 0.368(T5) - 0.000195(ENRL) + .008228(DENS) \\ - 3.221(IMTR) - 3.319(IM) - 44.85(PLUN)$$

Given the final equation, we can interpret its coefficient to make predictions on how best average 8th grade test scores can be explained – given data on GA school districts in 1996. The parameter estimates will be taken as partial derivatives of the multivariate regression. All else equal, a 1-point increase in 3rd grade math and reading test scores yields a 0.181-point increase in average 8th grade math and reading test scores. A1-point increase in 5th grade math and reading test scores yields a 0.368-point increase in average 8th grade math and reading test scores. A1-student increase in school district enrollment yields a 0.0002-point decrease in average 8th grade math and reading test scores. A population density increase of 1 unit (likely people per square-mile) results in a 0.008-point increase 8th grade scores. If a district is located in a metropolitan area, 8th grade test scores will decrease by 3.221 points. Likewise, a medium-enrollment district will experience a 3.319-point decrease in 8th grade scores. Finally, a 1-unit increase in the proportion of students in a district on free lunch results in a 44.85-point decrease in average 8th grade math and reading test scores.

Here, it is not practical to interpret the intercept as anything more than a global mean because there will never be a case when average 3rd grade math and reading test scores will be 0, and the same applies for all the other predictor. In order to find the predicted average 8th grade test score, any combination of T3, T5, ENRL, DENS, IMTR, IM, and PLUN values can simply be plugged into the equation.

At last, it is important to note that the variables included in the model were selected using a backward elimination procedure based on an Akaike information criterion or (AIC). This threshold for parameter inclusion/exclusion is fairly lenient compared to that of a Bayesian information criterion (BIC). Final regressions can be as inclusive or as exclusive as the threshold for parameter p-values allows. The effects of varying levels of "strictness" when searching for the best model can be seen in Appendix I

V: CONCLUSION

Given "gasd96.txt" dataset on Georgia school district characteristics and performance scores, a detailed regression analysis results in the below model, relying on the statistically significant predictors: average 3rd grade math and reading scores (T3), average 5th grade math and reading scores (T5), district enrollment ENRL, population density (DENS), metropolitan area (IMTR), medium-enrollment district (IM) and the proportion of students on free lunch (PLUN).

$$\widehat{T8} = 64.13 + 0.1810(T3) + 0.368(T5) - 0.000195(ENRL) + .008228(DENS) \\ - 3.221(IMTR) - 3.319(IM) - 44.85(PLUN)$$

While this model aims to predict average 8th grade test scores based on a host of contributing factors, it is important to recognize that there are several other factors that contribute to student achievement and test outcomes – including, class size, school type, teacher quality, etc. While this final model achieves nearly the same predictive ability as the full model (containing all 8 numerical and binary categorical variables), the magnitude of residuals (difference between observed and predicted values) reveals the noisiness of both the predictor variables and the predicted results. For instance, the model predicts that Appling County, given its observable characteristics included in the data set, will have average 8th grade scores of 96.96 (rounds to 97). But in 1996, the average 8th grade score in Appling County was 105. The model also predicts that Valdosta County will have a score of 88.18 (rounds to 88), when it's observed score was 84.

The inclusion of DENS, IMTR, and IM in the final model perhaps demonstrates that the size and location of a district does impact average 8th grade test scores. For instance, highly populated districts typically higher test scores, on average. Likewise, from this analysis we can claim that districts with a high proportion of students who are eligible for free lunch, usually students from lower-socioeconomic statuses, are likely to have lower average test scores. In addition, positive performance on 5th grade math and reading exams is likely to have a stronger impact on 8th grade scores rather than 3rd grade math and reading results.

Despite the limitations in data availability and prediction, this analysis is still valuable in the context of education policy. For most public schools in Georgia, 8th grade marks the end of middle school and the point of transition into high school. Thus, it is useful to understand the relative impacts of student, school, and district characteristics on the most common measure of student achievement – test scores.

APPENDIX I: SUPPLEMENTARY OUTPUT

Regression Output for Full Model:

$$T8_i = \beta_0 + \beta_1 T3_i + \beta_2 T5_i + \beta_3 \ln(ENRL)_i + \beta_4 \ln(DENS)_i + \beta_5 IMTR_i + \beta_6 IM_i + \beta_7 IL_i + \beta_8 PLUN_i + \varepsilon_i$$

Number of Observations Read	174
Number of Observations Used	174

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	8	43322	5415.27937	65.14	<.0001
Error	165	13716	83.12919		
Corrected Total	173	57039			

Root MSE	9.11752	**R-Square**	0.7595
Dependent Mean	95.20690	**Adj R-Sq**	0.7479
Coeff Var	9.57653		

Variable	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	64.09332	9.59803	6.68	<.0001
T5	0.37345	0.08797	4.25	<.0001
T3	0.17773	0.06415	2.77	0.0062
ENRL	-0.00014566	0.00010088	-1.44	0.1507
DENS	0.00786	0.00321	2.45	0.0153
IMTR	-2.68301	2.09735	-1.28	0.2026
IM	-3.86654	1.64831	-2.35	0.0202
IL	-3.04386	3.97789	-0.77	0.4452
PLUN	-44.92547	7.20196	-6.24	<.0001

Variable Elimination Based on Different Criterion:

1. Backwards Elimination Based on P-value of 0.01 (same as BIC Criterion results in R Code)

All variables left in the model are significant at the 0.01 level.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	42001	14000	158.28	<.0001
Error	170	15037	88.45496		
Corrected Total	173	57039			

Variable	**Parameter Estimate**	**Standard Error**	**Type II SS**	**F Value**	**Pr > F**
Intercept	57.52347	9.66352	3134.30493	35.43	<.0001
T5	0.33428	0.08862	1258.51458	14.23	0.0002
T3	0.22466	0.06478	1063.74673	12.03	0.0007
PLUN	-38.15086	7.06242	2581.21269	29.18	<.0001

2. Backwards Elimination Based on P-value of 0.05

All variables left in the model are significant at the 0.05 level.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	6	43053	7175.43559	85.68	<.0001
Error	167	13986	83.74813		
Corrected Total	173	57039			

Variable	**Parameter Estimate**	**Standard Error**	**Type II SS**	**F Value**	**Pr > F**
Intercept	61.63605	9.51017	3517.77314	42.00	<.0001
T5	0.37274	0.08802	1501.70088	17.93	<.0001

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
T3	0.18161	0.06424	669.23458	7.99	0.0053
ENRL	-0.00020508	0.00007763	584.43451	6.98	0.0090
DENS	0.00655	0.00301	396.83356	4.74	0.0309
IM	-3.57324	1.48224	486.70163	5.81	0.0170
PLUN	-41.68860	6.96136	3003.45877	35.86	<.0001

3. Backwards Elimination Based on AIC Criterion (less stringent compared to BIC)

Analysis of Variance				
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**
Model	7	43274	6181.93727	74.55
Error	166	13765	82.92163	
Corrected Total	173	57039		

Root MSE	9.10613
Dependent Mean	95.20690
R-Square	0.7587
Adj R-Sq	0.7485
AIC	952.52415
AICC	953.62171
SBC	801.79659

Parameter Estimates				
Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**
Intercept	1	64.132411	9.585908	6.69
T5	1	0.368555	0.087625	4.21
T3	1	0.181004	0.063927	2.83
ENRL	1	-0.000195	0.000077494	-2.52
DENS	1	0.008228	0.003165	2.60
IMTR	1	-3.221268	1.973406	-1.63
IM	1	-3.319109	1.483099	-2.24
PLUN	1	-44.847998	7.192249	-6.24

$$T8_i = \beta_0 + \beta_1 T3_i + \beta_2 T5_i + \beta_3 ENRL_i + \beta_4 DENS_i + \beta_5 IMTR_i + \beta_6 IM_i + \beta_7 PLUN_i + \varepsilon_i$$

In this paper, we choose this model to be our final regression because it is the most comprehensive according to AIC Backward Elimination. Final regressions can be as inclusive or as exclusive as the threshold for parameter p-values allows.

APPENDIX II: PROGRAM CODES

1. R CODE (PRIMARY)

```
# STAT 4360 Final Project
# Shreya Ganeshan
# 29 November 2016

# ----------------- Importing Libraries ------------------
library(MASS)
library(car) # for Durbin Watson Test and Variance Inflation Factor (VIF)
library(leaps) # for regsubsets
# ----------------- Reading in Dataset ------------------
# clear all previous variables
rm(list=ls())
data = read.table("gasd96.txt", header = T)
data
attach(data) # to call variables by name
detach(data)

# ----------------- Editing Variables ------------------
# Variables: COUNTY, T8, T3, T5, ENRL, DENS, IMTR, IM, IL, PLUN

# changing variable responses to binary categorical variables
NameIMTR = IMTR
NameIMTR[NameIMTR == 1] = "Metro Area"
NameIMTR[NameIMTR == 0] = "Non-Metro Area"
IMTR=NameIMTR
IMTR

NameIM = IM
NameIM[NameIM == 1] = "Medium Enrollment"
NameIM[NameIM == 0] = "Non-Medium Enrollment"
IM=NameIM
IM

NameIL = IL
NameIL[NameIL == 1] = "Large Enrollment"
NameIL[NameIL == 0] = "Non-Large Enrollment"
IL=NameIL
IL

# ----------------- Parsing Dataset ------------------
# dataset with only numeric variables
# removing OBS (column 1), COUNTY (column 2), IMTR (column 8), IM (column 9),
IL (column 10)
datanumeric = data[,(c(3,4,5,6,7,11))]
datanumeric

# dataset with only binary categorical variables
datacat = data[,(c(8,9,10))]
datacat

# -------------------------------------------------------
# SECTION 3: DATA SUMMARY
# ----------------- Summary Statistics: Numerical ------------------
```

```
#
summary(data)
stem(T8)
sd(T8)
sd(T3)
sd(T5)
sd(ENRL)
sd(DENS)
sd(PLUN)

# creating a function to calculate summary stats all at once for all
numerical variables
summarystats = function(x) {
  c(Mean=mean(x), Median=median(x), STDV=sd(x), Min=min(x), Max=max(x),
  Q1=quantile(x,(1/4),names=F), Q3=quantile(x,(3/4),names=F), IQR=IQR(x))
}
# creating a dataframe (vector) to store al the summary stats for all
numerical variables
stats = data.frame(summarystats(T8), summarystats(T3), summarystats(T5),
summarystats(ENRL),
                   summarystats(DENS), summarystats(PLUN))
stats
#transpose and round the stats dataframe/matrix
stats = t(stats)
stats = round(stats,3)
stats

# creating a new table (writing a csv file) to store summary stats of all
numeric vars
write.csv(stats, file = "finalsummarynumeric.csv", quote=F)

# ----------------- Summary Statistics: Categorical ------------------
# does not make sense to tabulate summary statistics, so use frequency tables
# though, if categorical variables were still coded as 0 or 1, then finding
the mean/median
  # could elicit how if there are more or less 0 or 1 values depending on how
close
  # mean/median values are to 0.5
# freqencies of categorical variables: IMTR, IM, IL
IMTRfreq = IMTR
IMTRfreq = table(IMTRfreq)
IMTRfreq # 0 = 124 and 1 = 50

IMfreq = IM
IMfreq = table(IMfreq)
IMfreq # 0 = 107 and 1 = 67

ILfreq = IL
ILfreq = table(ILfreq)
ILfreq # 0 = 155 and 1 = 19

# lisitng all binary categorical var in frequency table
freqtable = data.frame(IMTRfreq, IMfreq, ILfreq)
freqtable
# renaming columns
colnames(freqtable)[1] = "IMTR"
colnames(freqtable)[2] = "Frequency"
colnames(freqtable)[3] = "IM"
colnames(freqtable)[4] = "Frequency"
colnames(freqtable)[5] = "IL"
colnames(freqtable)[6] = "Frequency"

# creating a new table (writing a csv file) to store frequencies of all
categorical vars
write.csv(freqtable, file = "finalfreqcat.csv", quote=F)

# ----------------- Box Plots of Each Variable ------------------
# Numerical Variables
datanumeric
# standardizing variables (distance from respective means / respective SD)
datanumericstd = scale(datanumeric, center = T, scale = T)
datanumericstd
boxplot(datanumericstd, col = "grey", ylab = "Standardized Observations",
        xlab = "Numerical Variables", main = " Distribution of Standardized
Numerical Variables")

boxplot(T8, col="grey", ylab = "Score Values", main = "Summary of 8th Grade
Scores in all Districts")

# ----------------- Plots of T8 vs. Each Variable ------------------

plot(modelfinal)
par.plot = par(mfrow=c(2, 4))
plot(T8 ~ T3, xlab = 'T3', ylab = 'T8',main = 'T8 vs T3')
plot(T8 ~ T5, xlab = 'T5', ylab = 'T8', main = 'T8 vs T5')
plot(T8 ~ ENRL, xlab = 'ENRL', ylab = 'T8', main = 'T8 vs ENRL')
#plot(T8 ~ log(ENRL), xlab = 'log(ENRL)', ylab = 'T8', main = 'T8 vs
log(ENRL)')
plot(T8 ~ DENS, xlab = 'DENS', ylab = 'T8', main = 'T8 vs DENS')
#plot(T8 ~ log(DENS), xlab = 'log(DENS)', ylab = 'T8', main = 'T8 vs
log(DENS)')
plot(T8 ~ PLUN, xlab = 'PLUN', ylab = 'T8', main = 'T8 vs PLUN')
boxplot(T8 ~IMTR, xlab = 'IMTR', ylab = 'T8', main = 'T8 vs IMTR')
boxplot(T8 ~IM, xlab = 'IM', ylab = 'T8', main = 'T8 vs IM')
boxplot(T8 ~IL, xlab = 'IL', ylab = 'T8', main = 'T8 vs IL')

pairs(c(datanumeric,datacat)) # good for appendix

# par(mfrow=c(1,1))
barplot(IMTRfreq, col=c("grey", "brown"), ylim = c(0,125),
        main = "Number of School Districts in Metro Areas", names.arg
=c('Non-Metro', 'Metro'))
# no = 124, yes = 50

barplot(IMfreq, col=c("grey", "brown"), ylim = c(0,115),
        main = "Number of Medium-Enrollment Districts in GA", names.arg
=c('Medium', 'Non-Medium'))
# no = 107, yes = 67
# -------------------------------------------------------
# SECTION 4: ANALYSIS
# ----------------- Correlation of All Variables ------------------
# numerical
corrnum = round(cor(datanumeric),2)
corrnum
```

```
#writing this into a CSV table
write.csv(corrnum, file = "correlationmatrix.csv", quote=F)

# cateogrical
corrcat = round(cor(datacat),2)
corrcat
#writing this into a CSV table
write.csv(corrcat, file = "correlationmatrixcat.csv", quote=F)

# ----------------- Checking Assumptions ------------------
# normality fo T8
# no clear pattern in residuals
# independence/random sampling
# can also plot residual diagnostics (but can't do this without the model)

# ----------------- Plotting Normality of Response Variable -----------------
-
par(mfrow=c(1,1))
# plot density function
density = density(T8) # uniform density
plot(density, main = 'Density Distribution of T8')

# normal curve over distribution of T8
h = hist(T8, main = 'Distribution of T8', col = 'grey', breaks = 10, xlab =
'T8')
xfit<-seq(min(T8),max(T8),length=40)
yfit<-dnorm(xfit,mean=mean(T8),sd=sd(T8))
yfit <- yfit*diff(h$mids[1:2])*length(T8)
lines(xfit, yfit, col="blue", lwd=2)

shapiro.test(T8)
# fail to reject the null hypothesis that sample comes from a population
  # with normal distribution

# ----------------- Regression Model ------------------
# model 1 includes all numerical and binary categorical variables in the
model
model1 = lm(T8~ T3 + T5 + ENRL + DENS + (IMTR==1) + (IM==1) + (IL==1) + PLUN,
data = data)
summary(model1)
# make sure the categorical indicators are coded as 1 == TRUE in reg
vif(model1) # no coefficient has very high multicolinearity

# model 2 includes ln(DENS) and ln(ENRL)
model2 = lm(T8~ T3 + T5 + log(ENRL) + log(DENS) + (IMTR==1) + (IM==1) +
(IL==1) + PLUN, data = data)
summary(model2)

# MODEL SELECTION:
# backward selection - instructions for project
step1 = stepAIC(model1, direction = 'backward', data = data)
step1$anova
step2 = stepAIC(model1, direction = 'both', data = data)
step2$anova

# model3 is the one identified as the best by BACKWARD
# This is still not as good: T8 ~ T3 + T5 + log(ENRL) + log(DENS) + IMTR + IM
+ PLUN - IM
model3 = lm(T8~ T3 + T5 + ENRL + DENS + (IMTR==1) + (IM==1) + PLUN, data =
data)
summary(model3)

vif(model3) # no coefficient has very high multicolinearity
summary(influence.measures(model3)) # attempting to find influential points

par(mfrow=c(1,1))

# another way to check best model: use BIC criterion
all = regsubsets(T8~(T3 + T5 + DENS + ENRL + PLUN + IMTR + IM + IL),data
=data, nvmax = 10, really.big = T)
summary(all)
summary(all)$bic
plot(1:8, summary(all)$bic,type="b",pch=19,col="sienna")
subsets(all, statistic = "bic", main = 'Best Model to Predict T8')
# 3 VARIABLES = BEST MODEL but this is based on stricter p-value threshold
(0.01)

# ----------------- Final Regression Model ------------------
modelfinal = lm(T8~ T3 + T5 + ENRL + DENS + (IMTR) + (IM) + PLUN, data =
data)
summary(modelfinal)
anova = anova(modelfinal)
aov(modelfinal)
write.csv(anova, file = "anova.csv", quote=F)

# tried interaction term between T3 and T5 in the model:
# (T3:T5 + T3 + T5) and (T3:T5) but term was never significant

# ----------------- Regression Diagnostics ------------------
par(mfrow=c(2,2))
plot(modelfinal)

# ----------------- Cook's Distance/Influential Points ------------------

summary(influence.measures(modelfinal))
par(mfrow=c(1,1))
plot(cooks.distance(modelfinal))

# ----------------- Plotting Observed vs. Predicted Values ------------------
# regression residual values
residuals = (summary(modelfinal))$residuals
residuals

# predicted values from regression
predicted = predict(modelfinal)

# plot of observed T8 versus predicted T8
plot(T8~predict(modelfinal), main = 'Observed T8 vs Predicted T8 Values',
xlab = 'Predicted',
     ylab = 'Observed')

# abline(modelfinal) hard to do for all 8 predictors at once
# ----------------- Durbin-Watson Test ------------------
# tests for time dependence among residuals
dwt(residuals)
# ----------------- Chi-Squared Test ------------------
chisq.test(abs(residuals))

# ----------------- Breusch-Pagan Test ------------------
# tests for homoskedasticity or non-constant variance (again)
ncvTest(modelfinal)

# ----------------- Variance Inflation Factor ------------------
vif(modelfinal)

# ----------------- Model Comparison ------------------
# null model - intercept is expected value of T8 (mean)
# all rest beta parameters are 0
# meanT8 = as.numeric(mean(T8))
# meanT8rep = rep((meanT8rep), 174)
# meanT8rep = c(meanT8rep)
# meanT8rep
# is.numeric(meanT8rep)
#cbind(meanT8rep)

modelnull = lm(T8 ~ 1)
summary(modelnull)

# full model - all numerical and binary cateogrical
modelfull = model1
modelfull
summary(modelfull)
write.csv(fullsum, file = "fullmodelreg.csv", quote=F)


# final model - T3, T5, ENRL, DENS, IMTR IM, PLUN
modelfinal
summary(modelfinal)

#compare the two
anovacompare = anova(modelnull, modelfull, modelfinal)
write.csv(anovacompare, file = "modelcomparison.csv", quote=F)

# ----------------- Predicting Value ------------------
data
# predict T8 for APPLING County
64.13 + (0.181*106) + (0.3686*116) - (0.000195*3494) +(.008228*32.10) -
(3.221*0) - (3.319*1) - (44.85*0.57184) #105 actual and 96.69
# predict T8 for VALDOSTA County
64.13 + (0.181*97) + (0.3686*98) - (0.000195*7455) +(.008228*166.53) -
(3.221*0) - (3.319*1) - (44.85*0.58471) #84 actual and 88.18304
```

2. SAS CODE (SECONDARY)

```
*STAT 4360 FINAL PROJECT - Shreya Ganeshan;

dm 'log;clear;output;clear;';
options ps=50 ls=75 pageno=1;

data data;
      infile 'C:\Users\PWD193\Downloads\gasd96.txt' expandtabs firstobs=2;
      input obs COUNTY$ T8 T3 T5 ENRL DENS IMTR IM IL PLUN;
run;

proc print data = data;
run;

* DATA SUMMARY SECTION;

*summary statistics of all variables;
proc means data = data mean std min max;
      var T8 T3 T5 ENRL DENS IMTR IM IL PLUN;
      * these are all of the variables of interest;
      * don't need COUNTY and OBS in the regression analysis;
      * IQR is displayed in R code;
run;

* distribuion of all variables - the numerical variables are the only ones of
interest;
* we see that all numerical variables are fairly symmetrical and normally
distributed ;
* this is evident in the histograms;
proc univariate data = data;
      var T8 T3 T5 ENRL DENS IMTR IM IL PLUN;
      histogram;
run;

* frequency tables of all binary categorical variables;
* boolean values are recoded to reflect relevant variable interpretations in
R code;
proc freq data = data;
      tables IMTR IM IL;
run;


* plotting T8 against all numerical variables ;
* does not make sense to plot categorical variables on a scatter plot;
proc gplot data = data;
      plot T8*T5 T8*T3 T8*ENRL T8*DENS T8*PLUN;
run;

* plotting T8 against all categorical variables;
* need to sort data by each variable first to separate out the binary values;

* IMTR;
proc sort data = data out = IMTRdata;
      by IMTR;
run;
```

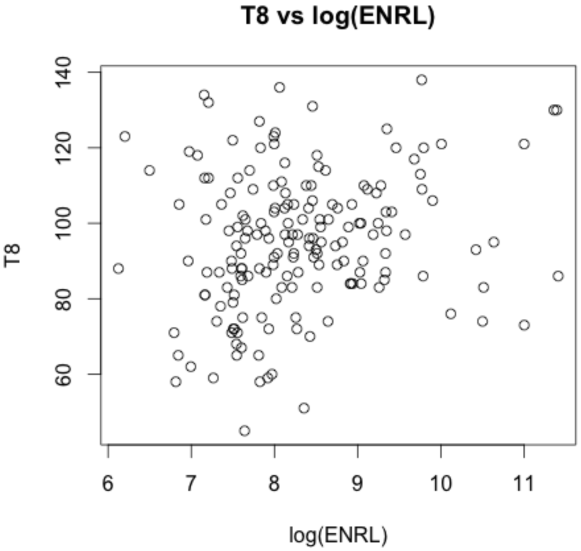
```
proc boxplot data = IMTRdata;
      plot T8*IMTR;
run;

*IM;
proc sort data = data out = IMdata;
      by IM;
run;
proc boxplot data=IMdata;
      plot T8*IM;
run;

*IL;
proc sort data = data out = ILdata;
      by IL;
run;
proc boxplot data = ILdata;
      plot T8*IL;
run;

* ANALYSIS SECTION;

* correlation matrix
* measures strength of linear assocation between T8 and all variables;
* don't have to separate by data type in SAS like you must do in R;
* can find correlation between numerical and categorical vars;
* p-values also displayed in SAS;
proc corr data = data;
      var T8 T5 T3 ENRL DENS IMTR IM IL PLUN;
run;

* regression on with all variables - excluding COUNTY and OBS;
proc reg data = data;
      model T8 = T5 T3 ENRL DENS IMTR IM IL PLUN;
run;
* all residual diagnostics can be seen use just the code above;
* this output also gives R2 value;

* model/variable selection;
* backward selection based on p-value of 0.01 - best model is: T5 T3 PLUN;
proc reg data = data;
      model T8 = T5 T3 ENRL DENS IMTR IM IL PLUN / selection=backward
slstay=.01;
run;

* backward selection based on p-value of 0.05 - best model is: T5 T3 ENRL
DENS IM PLUN;
proc reg data = data;
      model T8 = T5 T3 ENRL DENS IMTR IM IL PLUN / selection=backward
slstay=.05;
run;
* the smaller the p-value threshold the narrower the model gets;

* backward selection based on AIC - best model is: T5 T3 ENRL DENS IMTR IM
PLUN;
* stepAIC is more reliable than p-value;
proc glmselect data = data;
```

```
      model T8 = T5 T3 ENRL DENS IMTR IM IL PLUN /
selection=backward(choose=AIC);
run;

* regression with bet predictors;
* final model;
proc reg data = data;
      model T8 = T5 T3 ENRL DENS IMTR IM PLUN;
run;

* residual diagnostics - checking assumptions;
* normality;
* constant variance or homoskedasticity;
* indepenence;
* outliers and influential points;
proc reg data = data;
      model T8 = T5 T3 ENRL DENS IMTR IM PLUN / r p dw;
      output out = data2 r = residuals p = predicted;
      * cook's distance calulated along with regression output;
run;

* check to see that data2 was actually stored with predicted and residual
values;
* HAVE to have predicted and residual values to plot residual diagnostic
plots;
proc print data = data2;
run;

* residual diagnostic plots;
goptions reset=all;

* observed vs predcted values;
title1'Observed vs Predicted Values';
proc gplot data = data2;
      plot T8*predicted / haxis=axis1 vaxis=axis2 frame grid;
      axis1 label = (a=90 'Predicted');
      axis2 label = ('Observed');
run;

* residual vs fitted alues;
title1'Residual vs Fitted Values';
proc gplot data = data2;
      plot residuals*predicted / haxis=axis1 vaxis=axis2 frame grid;
      axis1 label = (a=90 'Fitted');
      axis2 label = ('Residuals');
run;

* qq-plot;
title1'QQ Plot';
proc univariate data = data2 normal;
      * "normal" options checks to see if resid. are normalyly dist;
      var residuals;
      qqplot residuals / normal(sigm = est mu = est) square;
run;
* residuals are fairly on normal line;

* durbin-watson test - to see if residuals are time dependent;
```

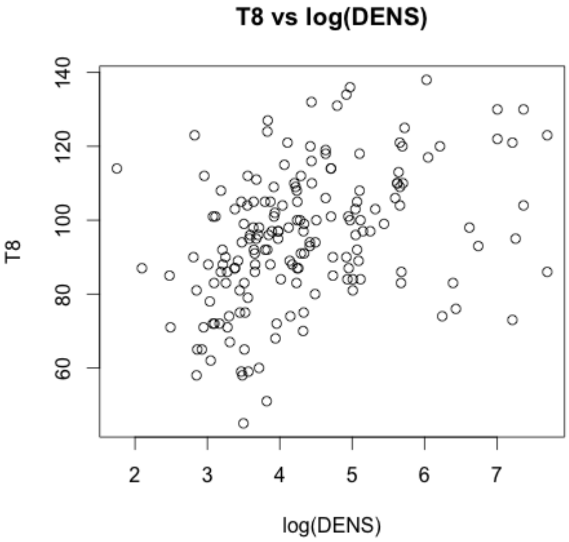
```
proc reg data = data;
      * can use regular "data" data set now;
      model T8 = T5 T3 ENRL DENS IMTR IM PLUN / dw;
      * will get more info in output if you add in "r, p, dwprob" options;
run;

* white test - to see if there is no homoskedasticity;
* null hypothesis = homoskedasticity
* bc of a high p-value we fail to reject the null hypothesis;
proc reg data = data;
      model T8 = T5 T3 ENRL DENS IMTR IM PLUN / spec;
run;

* model comparisons;
* null - basically, the expected value of T8;
proc univariate data = data;
      var T8;
run;
*full - all variables in the dataset of interest;
* same code from above;
proc reg data = data;
      model T8 = T5 T3 ENRL DENS IMTR IM IL PLUN;
run;

*final - only variables: T3, T5, ENRL, DENS, IMTR, IM, PLUN;
proc reg data = data;
      model T8 = T5 T3 ENRL DENS IMTR IM PLUN;
run;

* model comparison tables were just done by hand for the report, only using
R2 values;
* R2 values are most relevant to the reader;
```

APPENDIX III:

Plots of T8 vs ln(ENRL) and T8 vs ln(DENS) to demonstrate relatively positive linear association after logarithmic transformation.



T8 vs log(ENRL)



T8 vs log(DENS)