Horace Liu and Joseph Rose

[redacted]

Professor Perez-Urdiales

ECO 321: Econometrics

11 December 2020

An Investigation into Nonlinearities, Interactions, and Instrumental Variables Affecting Factors Determining Average SAT Scores in Public New York City High Schools

1. Introduction

A previous analysis on the effect that student population size and external factors had on the respective schools' average SAT scores within public New York City high schools in 2015 concluded that schools with larger population sizes had higher average SAT scores. In addition, externalities such as students' ethnicities, number of students in poverty, and number of students in ELL programs showed stronger correlations with the SAT scores than population size. When generating visualizations, it visually seemed that the number of students in poverty and the number of students in ELL programs were more closely bunched together along a line, noting that it was entirely possible that the number of students in ELL programs and the number of students in poverty at each school had a more significant impact on the average SAT score at that school. On the visualizations, it was also noted that the points were bunched together in a nonlinear relationship between average SAT score, and number of ELL students or the number of students in poverty at that school. Furthermore, in the article, "Race, Poverty and SAT Scores: Modeling the Influences of Family Income on Black and White High School Students' SAT Performance" by Ezekiel J. Dixon-Roman et al., it was proven that family income has a positive

concave relationship with students' SAT scores (Dixon-Roman et al., 2013). This indicates that there are nonlinear relationships between family income and SAT scores, which can be extrapolated into a nonlinear relationship between students in poverty and their SAT scores. Thus, a limitation is raised from the results and conclusions of the previous investigation. As nonlinearities were ignored, as well as possible interactions between variables, there may have been greater errors in the findings, which will be explored. Therefore, the question becomes: to what extent do nonlinearities, interactions between variables, and instrumental variables influence factors that affect average SAT scores in public New York City high schools?

By investigating linear-log and quadratic regressions, possible interactions between certain variables that may be correlated, and instrumental variables through the use of OLS and TSLS, perhaps a model can be found that fits better than the previous investigation to explain factors which influence each schools' average SAT scores.

2. Econometric Model Outline

Foremost, in order to generate the OLS equation describing potential factors, nonlinearities, and interactions affecting average SAT scores, an OLS model has to be explored. The Ordinary Least Squares (OLS) model is an econometric model that computes the minimization of the average of the differences in each of the points of the dependent variable on the graph and their predicted values on the residual line. It is useful in predicting the parameters that represent a specific linear model. As for interaction variables, they are two or more independent variables that affect each other or another independent variable. If two interaction variables are continuous-continuous variables, then the coefficient of a third independent variable is affected by a unit increase in two interaction variables.

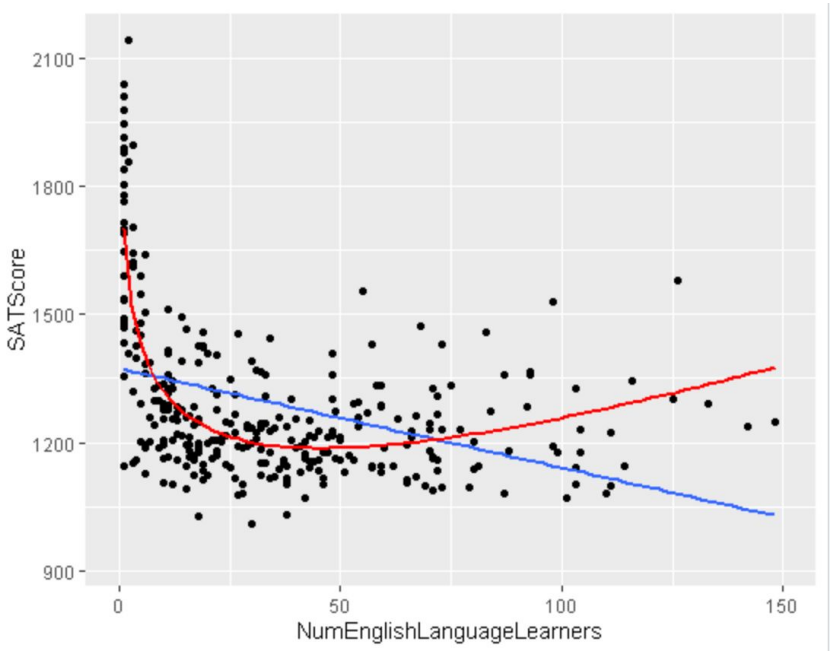
As expressed before, previous investigations into factors affecting average SAT scores in public New York City high schools ignored nonlinear factors, as well as interactions between certain variables. To simplify this investigation, separate SAT scores for each section (Reading, Writing, and Math) were combined for each school for an overall look into average SAT score at each school. Furthermore, as before, the factors that are being investigated in terms of affecting average SAT scores is the student population size, the number of students in an ELL program, the number of students in poverty, and the number of students of different ethnicities within each school. An additional variable, students of mixed or multiple ethnicities, was left out to avoid the dummy variable trap. Nonlinearities and interaction variables were added as a plausible explanation to a better fit as the factors affecting average SAT score.

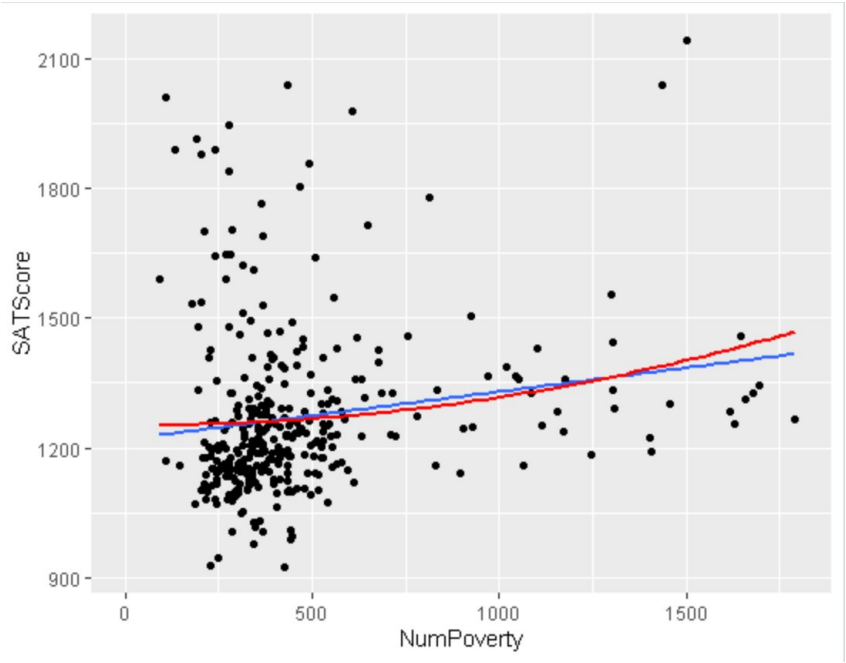
$$\text{SAT Score} = \beta_0 + \beta_1(\text{Total Enrollment}) + \beta_2(\text{English Language Learners}) + \beta_3(\log(\text{English Language Learners})) + \beta_4(\text{Students in Poverty}) + \beta_5(\text{Students in Poverty})^2 + \beta_6(\text{Asian}) + \beta_7(\text{White}) + \beta_8(\text{Black}) + \beta_9(\text{Hispanic}) + \beta_{10}(\text{English Language Learners} * \text{Students in Poverty}) + u_i$$

3. Econometric Model Relations

When visualizations of each variable were generated as the independent variable, with the average SAT score as the dependent variable, it was noted visually that two variables in specific, number of students in an ELL program and the number of students in poverty, had a rough nonlinear correlation with the average SAT score. In particular, the number of students in an ELL program as the independent variable displayed a hyperbolic curve with logarithmic tendencies. This would make sense as English Reading and Writing is a large part of SAT

exams, and if there are more ELL students in a school taking the SAT, the average would be lower than schools with less ELL students as lower English proficiency would mean lower SAT scores. As some schools had no ELL students, those values had to be changed to 1, as log(0) would be undefined. As for the number of students in poverty as the independent variable, visually, there seemed to almost be a quadratic relationship with average SAT scores. In theory, it should be the opposite, as students with more money would be more likely to have more access to educational resources, but this could be a direct result of the fact that public education is free in New York City between ages from 5 to 21 (Infante-Green, 2016). Lastly, an interaction variable was added between the number of students in an ELL program and the number of students in poverty as it is entirely possible that students in poverty would have less access to education resources, and therefore would possibly have lower English proficiency.





When the nonlinear terms are graphed individually against the SAT score, it seems that visually for the plot where the Number of ELL students are the independent variable, the nonlinear curve fits the data much better than a linear curve. However, for the graph where the Number of Students in Poverty is the independent variable, visually, the nonlinear curve might not show much of a difference but it is significant when considering the impact of the nonlinear term in the overall model, which will be explained in detail.

4. Results

Under heteroskedastic robust standard errors, the data was generated for the model for the estimated coefficient, standard error, test statistic, and the significance at $\alpha = 0.01$.

Regressor	Description	Estimated Coefficient	Standard Errors	T-Statistic	P-Value
Intercept	The intercept of the	1491.1	32.133	46.4036	$< 2.2*10^{-16}$

	regression model				
Total Enrollment (Dummy)	Dummy variable of Total Enrollment where enrollment size > 700 = 1, otherwise, = 0	67.165	22.140	3.0337	0.0025923
English Language Learners	Number of students in an ELL program per school	0.20706	0.10755	1.9252	0.0549913
log(English Language Learners)	The natural log of the number of students in an ELL program per school	- 101.4	9.5095	- 10.6632	$< 2.2*10^{-16}$
Students in Poverty	Number of students in poverty per school	0.16516	0.082783	1.9951	0.0467926
(Students in Poverty)^2	The square of the number of students in poverty per school	$- 8.2982*10^{-5}$	$2.1377*10^{-5}$	- 3.8818	0.0001234
Asian	Number of asian students per school	0.28274	0.053325	5.3023	$2.006*10^{-7}$
White	Number of white students per school	0.087222	0.042882	2.0340	0.0426894
Black	Number of black	- 0.089496	0.055533	- 1.6116	0.1079353

	students per school				
Hispanic	Number of hispanic students per school	0.097414	0.062901	1.5487	0.1223409
(English Language Learners) * (Students in Poverty)	The interaction term between ELL students and students in poverty	$-5.945*10^{-7}$	$4.8339*10^{-5}$	- 0.0123	0.9901942

```
Residual standard error: 98.1 on 358 degrees of freedom
  (56 observations deleted due to missingness)
Multiple R-squared:  0.7552,    Adjusted R-squared:  0.7484
F-statistic: 110.5 on 10 and 358 DF,  p-value: < 2.2e-16
```

After analyzing the summary of the regression model, the multiple R-squared is 0.7552, and after correcting for the number of regressors, the adjusted R-squared is at 0.7484. This would mean that the model relatively accurately explains the dependent variable, SAT score, and that the data is relatively close to fitting the regression model.

Given the null hypothesis that all the nonlinear terms as well as the interaction term is equal to 0 (log(English Language Learners) = 0, $(\text{Students in Poverty})^2 = 0$ and (Num of ELL*Students in Poverty) = 0), at the 1% significance level, the F-statistic is quite large at 54.541, and the P value is quite small at $2.2*10^{-16}$. This would mean that the null hypothesis would be rejected, signifying that the nonlinear terms as well as the interaction term have influence on the overall model having a better fit than before. When looking at each of the nonlinear and interaction variables alone, it is noted that for log(English Language Learners), given the null hypothesis that the variable is equal to 0 at the 1% significance level and the absolute value test statistic, is relatively large at 10.6632 while the P value is very small at

$2.2*10^{-16}$. This means that the null hypothesis is rejected and that the nonlinear variable, log(English Language Learners), holds significance to the model. Furthermore, for the other nonlinear variable, $(\text{Students in Poverty})^2$, it is very similar, where given the null hypothesis that the variable is equal to 0 at the 1% significance level, the absolute value for the test statistic is large at 3.8818 and the P value small is 0.0001234, signifying that the null would be rejected and that this variable holds significance to the overall model despite visually not looking like there is much significance for the quadratic curve when compared against the linear curve. However, for the interaction term, given the null hypothesis that it is equal to 0 at 1% significance, the absolute value of the test statistic is small at 0.0123 and the P value is large at 0.9901942; this shows that the null hypothesis would not be rejected. Hence, there is a likelihood that the interaction term is equal to 0 at the 1% significance level, and that it is not very relevant to the model.

The interaction term's ineffectiveness is further exemplified when comparing the goodness of fit between the nonlinear model without the interaction term and the nonlinear model with the interaction term. For the nonlinear model with the interaction term, the adjusted R-squared is 0.7484, but when the interaction term is removed, the adjusted R-squared actually increased (albeit marginally) to 0.7491. The interaction term is therefore able to be determined as not very significant and can be removed from the nonlinear model in the future. It seems that the number of students in an ELL program and the number of students in poverty are not correlated as initially thought, thereby making the interaction term not very useful in the overall model.

When the initial nonlinear model, including the interaction term, is compared to previous linear models, there is evidently a significant increase in the adjusted R-squared value, signifying that there was a significant increase in the goodness of fit for the model. The previous linear model is shown below.

$$\text{SAT Score} = \beta_0 + \beta_1(\text{Total Enrollment}) + \beta_2(\text{English Language Learners}) + \beta_3(\text{Students in Poverty}) + \beta_4(\text{Asian}) + \beta_5(\text{White}) + \beta_6(\text{Black}) + \beta_7(\text{Hispanic}) + u_i$$

The adjusted R-squared for the linear model was 0.5071, and when compared to the nonlinear model with the interaction term at 0.7484, it is evident that a 0.24 increase had taken place, which, in terms of adjusted R-squared values, is a significant increase. This also helps support the fact that the nonlinear model is a much better fit than the linear model, and that including nonlinear components is vital to generating a better model for the dependent variable.

Given the results of this analysis, it is evident that one main factor within the nonlinear model is that schools with more students in an ELL program would have a relatively lower average SAT score. A new policy could be derived from that where New York City should focus more of its efforts in improving students' English proficiency, as that seems to have a large impact on average SAT scores of public highschools in New York City in 2015. Perhaps there could be free private tutoring and extra learning opportunities either after or before school to provide students in ELL programs with more opportunities to improve their language skills.

5. Instrumental Variables

Instrumental variables avoid any biases in a regression model from breaking the first Least Squares Assumptions by separating the independent variable X_i into correlated and uncorrelated components with the error term u, which leads to estimating β_1, the slope of the regression line. In general, there are three main biases that are threats to a regression being valid. First, the omitted variable bias explains a variable that is correlated with X but is not in the regression model. Second, the simultaneous causality bias is where the independent variable

affects the dependent variable and vice-versa. Third, there can be errors in the independent variable. All of these problems lead to $E(u|X) \neq 0$, which violates the first Least Squares Assumption. However, by using instrumental variables, the independent variable can be uncorrelated with u but correlated with Z. There are two types of variables that affect a regression model: endogenous and exogenous. Endogenous variables are ones that are correlated with the error term and exogenous variables are ones that are not correlated with the error term. Therefore, the independent variable is endogenous and the instrumental variable is exogenous. Furthermore, in order for an instrumental variable to be valid, it must be correlated with X, known as instrument relevance, and uncorrelated with u, known as instrument exogeneity. A well-known method in incorporating an instrumental variable to a regression model is the Two Stage Least Squares method, in which there are two stages. First, separate the X that is uncorrelated with u: $X_i = \pi_0 + \pi_1 Z_i + v_i$. Second, compute the estimator of X, $\hat{X}_i = \hat{X}_i + \hat{\pi}_1 Z_i$ to put into the original regression model, which leads to $Y_i = \beta_0 + \beta_1 \hat{X}_i + u$.

In terms of the regression model used for this analysis, there seems to be a viable assumption that the model suffers from an endogeneity bias. In particular, the observation of the Number of Students in Poverty variable, when used in the model as a regressor to determine the SAT scores, seems to suffer from omitted variable bias or confounding variables which would threaten internal validity. As the visualization of the regressor has been shown, there seems to be a minor quadratic curve to the data, meaning that as the number of students in poverty increases, there is a minor tendency for the average SAT score at the school to be higher. Logically, that shouldn't make sense, as students with more money would tend to have more and better access to educational materials, including private tutoring. It is possible that this is a result of education being free in New York, however, an instrumental variable could be implemented to test this.

An example of an instrumental variable that can be used to test for endogeneity bias is a scholarship or financial aid plan for students in poverty. This scholarship would be distributed randomly to students in poverty, instead of a scholarship by merit. Therefore, there would be no correlation between the instrumental variable and the error term, following the first assumption of the Least Squares Assumptions. Furthermore, having more financial flexibility can allow students to purchase laptops or textbooks for their classes and have more access to studying tools or private tutors. Therefore, the scholarships will be correlated with X, in this case the Students-in-Poverty variable, but not the error term u. To actually implement this, the part of Students-in-Poverty variable that is uncorrelated with u would be regressed on the scholarships variable using OLS to get the predicted value. SAT score would then be regressed on the result of the previous variable to get the TSLS estimator, which would be a consistent estimator of the original regressor. By correcting for endogeneity bias through the scholarship instrumental variable, it would help greatly in being able to determine if the number of students in poverty is a variable that would actually affect the average SAT score at a school.

6. Conclusion

Ultimately, the nonlinear regression model with interaction variables helped better define average SAT scores in public New York City highschools in 2015. The extent to which nonlinearities, interactions between variables, and instrumental variables influence factors that affect average SAT scores in public New York City high schools is quite large, with a 0.27 difference in adjusted R-squared values. When specifically looking at the interaction variables of two independent variables of the model, however, results showed that it was not useful in making the model better fitted because the test statistic became very small and the P value

became very large; thus, the interaction variable was not significant. Moreover, the regression model can be further improved upon by adding an instrumental variable (randomly distributed scholarships as an example) which controls for omitted variable bias and confounding variables as threats to internal validity for the students in poverty regressor within the nonlinear model. However, despite limitations to the model proposed in the analysis, the extrapolated economic policy from the results determine that the government can increase attention to students attending ELL programs by hiring more tutors or teachers who are fluent in speaking English, as well as increasing the number of hours for those courses. These actions allow students to be more aware, flexible and confident when approaching SAT problems which test heavily on English.

Works Cited

Dixon-Roman, E. J., Everson, H., & Mcardle, J. J. (2013, May). (PDF) Race, Poverty and SAT Scores: Modeling the Influences of Family Income on Black and White High School Students' SAT Performance. Retrieved October 29, 2020, from https://www.researchgate.net/publication/280232788_Race_Poverty_and_SAT_Scores_Modeling_the_Influences_of_Family_Income_on_Black_and_White_High_School_Students'_SAT_Performance

Infante-Green, A. (2016, May). Guidance Relating to the Right of Individuals Over Compulsory School Age to Attend High School. Retrieved December 6, 2020, from http://www.nysed.gov/common/nysed/files/memo_aig_school_age.pdf