A Regression Study on the Impact of Socioeconomic Factors on Student Performance

Kedar Bhingarde.

Introduction

This study seeks to determine if a relationship exists between student performance and a set of socioeconomic factors. These factors include class size, school expenditures, and family background. The outcome of studies such as this may help identify which factors policy makers should consider when working on improving educational standards.

Educational reform remains one of the most hotly debated national issues. There is recurring evidence that the American students are performing poorly in reading, mathematics and science when compared to their peers in the other countries. In 2013, the Organization for Economic Cooperation and Development (OECD) published the result of a comparative study of mathematics and science skills among 15-years old students in more than 60 nations and school systems. In mathematics, the U.S. ranked 26th. The rank was 21st in science.

There are almost as many reasons for this apparent failure of our public education as there are pundits who are willing to discuss it on national television. Poor teacher preparation, intransigence of teachers unions, poor parenting, and insufficient funding are some of the reasons given for this problem. At the same time, there is contrary view, documented by the American Federation of Teachers, which argue that the lack of early childhood education and poorly targeted resources are some of the contributing factors.

At the federal level, some efforts have been made to reform public education and improve student performance. Most notable were President George W. Bush's 2002 "No Child Left Behind" and President Barrack Obama's 2009 "Race to the Top". The former law required states to give students in grades 3-8 an annual test in reading and mathematics. Among other things, "Race to the Top" rewards states of implementing performance based standards for teachers and principals. Unfortunately, reforms such as these have yet to produce the intended results, which is to raise the proficiency level of high school graduates.

This study does not seek to explain or justify any factors believed to contribute to student's performance. Rather, it simply attempts to investigate **if a relationship exists** between student performance and a set of socio-economic factors often mentioned in the press as contributing to student performance. The method of analysis is a multiple regression.

1. PISA 2012 Results: www.oecd.org/pisa/keyfindings/pisa-2012-results.htm {retrieved on December 6, 2013}. The OECD study is conducted every three years.

Data and Methodology:

Empirical data for this study were obtained from 224 school districts in one of state in America. The response variable(Y) is student performance, and is measured by the mean score on the exam. The following four- variable multiple regression models are specified.

$Y=\beta0+\beta1X1+\beta2X2+\beta3X3+\beta4X4+\epsilon i$

Where,

X1=Student-to-teacher ratio (STR)

X2=Average teacher's salary (TSAL)

X3=Median household income (INC)

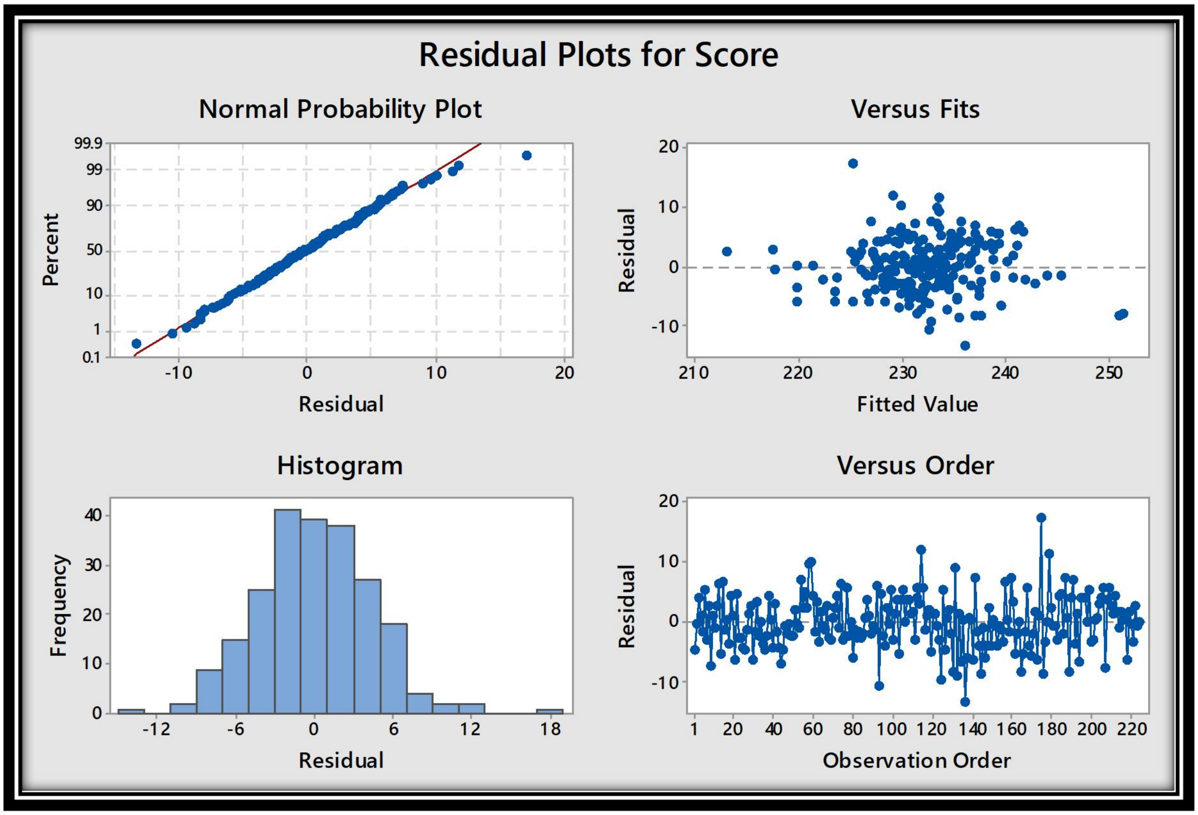
X4=Percentage of single family household (SGL)

Pursuant to the purpose of this study, the regression null hypothesis is as follows:

H0: $\beta1= \beta2= \beta3= \beta4=0$ **[There is no regression relationship]**

The main objective of this project is to study that socioeconomically which factors are to affected on the study on the students performance.

Data Analysis

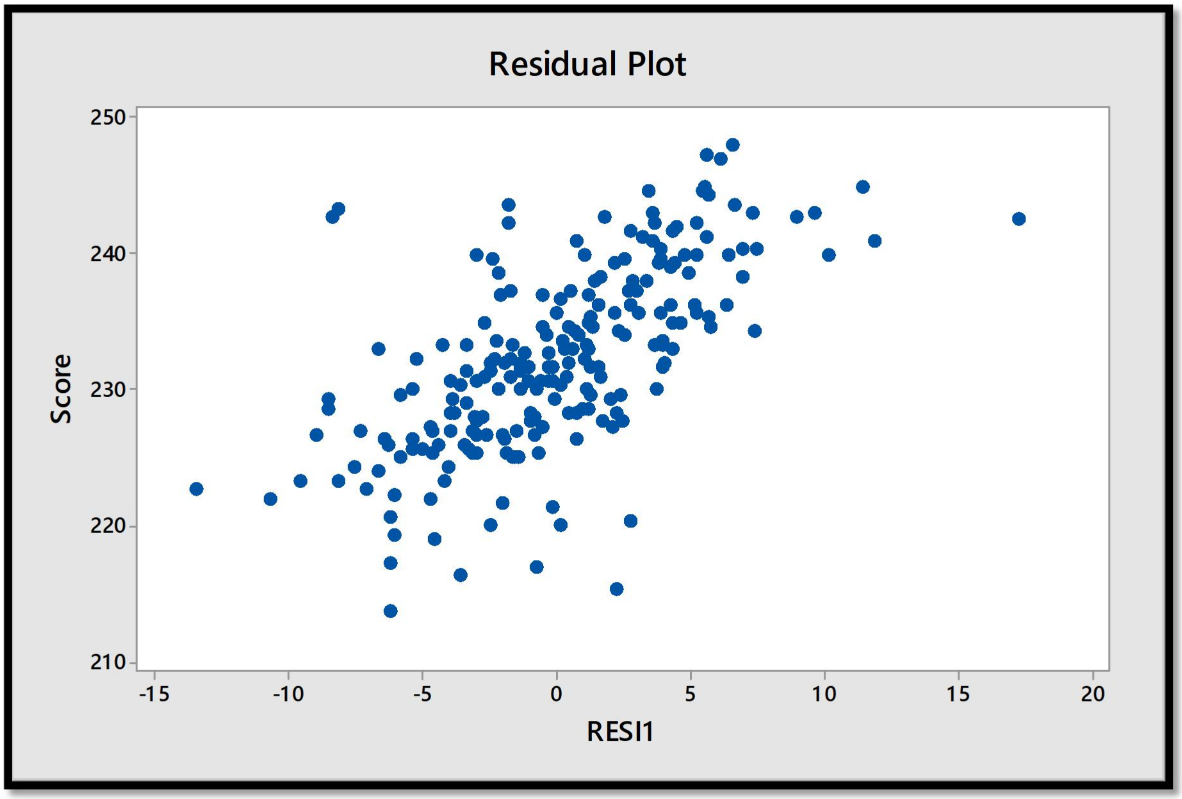


As here we see that in Normal Probability Plot the data is hugging the diagonal. Also, in Histogram we see that the all bars are closed to each other and normality is follow in Normal probability plot. As we see in Versus Fits plot the all points are merging in center which means that the heteroskedasticity is present. In Versus Order plot all points are lies in specific range(-10,10) which means (residual fluctuate is more or less in the random fashion inside band) then there are no obvious model effect.

Test for Heteroskedasticity

Heteroskedasticity, which is the problem of unequal error variance, is typically encountered with cross-sectional data. When Heteroskedasticity is present, the regression coefficients are no longer efficient. Loss of efficiency is because the standard error and confidence intervals are to narrow, giving a false sense of precision.

By using the residual plot we can check for the presence of Heteroskedasticity. If Heteroskedasticity is exists, the plot will tend to spread out as the value of x(or ŷ) increases. This means that as the value of x become larger, there is increasing uncertainty associated with the response of y. Figure1 shows the residual plot.



While the scatter of residual appear more robust in the mid-section of the graph than at the beginning, there is no defective indication that heteroscedasticity is present.

R-Output:

```
> y=scan("clipboard")

> x1=scan("clipboard")

> x2=scan("clipboard")

> x3=scan("clipboard")

> x4=scan("clipboard")

> regmodel=lm(y~x1+x2+x3+x4)

Call:

lm(formula = y ~ x1 + x2 + x3 + x4)

Coefficients:
```

Intercept	x1	x2	x3	x4
232.71377	-0.49451	-0.01294	0.26619	-0.89258

In the absence of any significant residual problems, the predication model is specified as follows:

The predicted model is:

$\hat{Y}=232.71377-0.49451x1-0.01294x2+0.26619x3-0.89258x4$

After predicting the model we see that the intercept of the model is 232.71377. And the explanatory variables are decreased in x1(student-teacher's ratio), x2(Teacher's salary), and x4(% of single family household) and in x3(Household income) which is to be increased in model.

> summary(regmodel)

Call:

lm(formula = y ~ x1 + x2 + x3 + x4)

Residuals:

Min	1Q	Median	3Q	Max
-13.3646	-2.9357	-0.1016	2.8141	17.2284

Coefficients:

	Estimate	Std.Error	t-value	Pr(>|t|)
Intercept	232.71377	3.47093	67.046	0.0000
x1	-0.49451	0.13476	-3.669	0.000305
x2	**-0.01294**	**0.07684**	**-0.168**	**0.86638**
x3	0.26619	0.03548	7.503	0.0000
x4	-0.89258	0.17899	-4.987	0.0000
Residual Standard error			4.436	
Multiple R-squared			0.5793	
Adjusted R-squared			0.5716	
F-statistics			75.38	
p-value			0.0000	

Regression results are summarized in above Table. The F statistic of 75.38(p-value=0.00) indicates that the regression, as a whole, is statistically significant. The coefficient of determination (r^2) suggests that about 57 percent of the variation in student performance is explained by four explanatory variable combined.

The test of significance for the independent variables- measured by the t statistics –shows that with the expectation of teachers salary (x2), all the explanatory variables are statistically significant. In other words, they contribute meaningful information in the prediction of student performance. The signs of coefficients for the independent variables are consistent with prior expectation with the exception of x2, which is not significant. Taking the value of the coefficient for x4 as an example, it suggest that the average student score decreases by about 0.89 for every percentage increase in single family household.

In the x2 the estimate (-0.01294) and its t-value (-0.168) corresponding its p-value (0.86638) which is not statistically significant.

ANOVA and Lack of Fit Test and PRESS Statistic:

Analysis of Variance:

Source	DF	SS	MS	F	P
Rgression	4	5933.8	1438.4	75.38	0.000
Residual Error	219	4309.6	19.7		
Total	223	10243.4			

S	4.436
R-Sq	57.90%
R-Sq(adj)	57.20%
PRESS	45.5084
R-Sq(pred)	55.57%

We know that the predicted residual sum of square (**PRESS**) statistics is a form of cross-validation used in regression analysis to provide a summary measure of the fit of a model to a sample of observations that were not themselves used to estimate the model.

Lack of test is one of the important tests to judge the efficacy of the model.

R-sq is 0.57 in this case where as R-Sq(adj) is 57.20%. This PRESS Statistics is comes around 45.51. Usually the lower the PRESS statistics better the model. However, considering the variability in the data this PRESS Statistics is quite good. Also, our predictive capacity of the model is 0.5557(55.57%) accurate our predictive model is.

Multicollinearity

Often, when collinearity exists, the sign of the regression coefficient is the opposite of what logic suggest. Also, the t-value may not be significant even when the F is significant. To confirm that the lack of statistical significance and illogical sign associated with x2(teacher's salary) is not due to multicollinearity between the independent variables are examined.

The correlation of half matrix is presented in table:

```
> d=data.frame(x1,x2,x3,x4)

> cor(d)
```

Table: **Correlation Matrix of Independent Variables**

	x1	x2	x3	x4
x1	1.0000	0.7729	-0.1635	0.2829
x2	0.7729	1.0000	0.5504	-0.2429
x3	-0.1635	0.5504	1.0000	-0.6178
x4	0.2829	-0.2429	-0.6178	1.0000

The correlation coefficient between teacher's salary(x2) and household income(x3) is 0.55, which may be considered high. Even larger is the correlation coefficient between household income(x3) and percentage of single family household(x4), which is about -0.61. This is however a casual observation.

A more rigorous analysis of multicollinearity is based on the calculation of the variance inflation factor (VIF) for each of the variable, defined as follows:

$\mathbf{VIF=1/(1-ri^{2})}$

The VIF value is obtained from running the following regressions:

$x1 = \alpha0 + \alpha1x2 + \alpha2x3 + \alpha3x4$

$x2 = \alpha0 + \alpha1x1 + \alpha2x3 + \alpha3x4$

$x3 = \alpha0 + \alpha1x1 + \alpha2x2 + \alpha3x4$

$x4 = \alpha0 + \alpha1x1 + \alpha2x2 + \alpha3x3$

The result of variation inflation factor – for each of the independent variables-are summarized in Table

Table: Variance Inflation Factors (VIF)

x1	1.12170499
x2	1.51331719
x3	2.24719101
x4	1.73100225

A high VIF means that the variance (and therefore standard error) of the regression coefficient is inflated, so that the corresponding t-value is less than it should be. A helpful rule of thumb is that collinearity exists if VIF>5. As Table shows, the highest VIF is 2.24, meaning that variable related to x3(household income) is only 2.24 times what it should be if collinearity did not exists. Finally we conclude that collinearity is probably not a problem in the model

Conclusions

- This study examined the relationship between student performance in English School and a set of socioeconomic factors. Student performance is measured by the mean score on the exam.

- Results shows that student-teacher ratio, teacher's salary, household income and the percentage of single family households as a whole affect student performance.

- Altogether, they account for about 57 percent of changes in the mean student assessment score. More specifically, the student performance rises with the household income, on average.

- There is also evidence that performance is restrict by a high student-teacher's ratio as well as by high incidence of single family households.

Limitation And Improvements

- **Limitations:**

It may also be helpful to determine if, after accounting for all the above factors, there is still a difference in performance between the various.

This latter inquiry can be pursued as a comparative study.

- **Improvements:**

An important improvement to this study is an investigation of how the study time spent by students outside of school and the level of education of the students parents might affect performance.