**Determinants of test scores for 5th grade students: divorce, extracurricular activities or school type?**

Hao Wu

May 11, 2021

**Abstract**

This project raises the question of determinants of test scores for 5th grade students. Classify the data set and eliminate missing values and invalid variables. The models and methods used are linear regression, least squares and fixed-effects models. After that, the data is grouped for regression and the results are as follows. Divorced and bad campus environment scores for students There are negative effects. Students in private schools have better scores than public schools, and participation in extracurricular activities has a positive effect on scores. Watching TV does not affect students' scores. The students' attitudes towards the crimes on campus were indifferent.

**1.Introduction**

The strength of a country is often multifaceted. It may be military, economy, or diplomacy; However, education is an indispensable part of a country’s strength. A country relies on the youth. Youth often face many problems. We need to find out those problems that really affect them.

Education is a part of a child’s growth into a youth, in which academic performance can also reflect the current situation of the student to a certain extent. The object of our research is 5th grade students. We will use the essence of data analysis to understand that What are the determinants of test scores for 5th grade students. In this research, we need to find out the decisive factors that affect children's grade, In the analysis process, it is necessary to exclude the useless information of the data set given by the nation center for education statistics. Our research mainly starts from the following questions whether school type (private versus public) impacts test scores? Does participation in extracurricular activities impact learning? and Are children negatively impacted by divorce? At the same time, we need to control the influence of race, region, and gender on our data.

**2. Literature Review**

According to the literature, there is a potential relationship between a student's performance and whether it is a single-parent family, as well as family income and community income. In low-income, high-poor communities, the percentage of households headed by women is higher, and schools with a lower percentage of adults with university degrees tend to have lower average test scores. And schools located in low-income areas tend to have fewer experienced teachers, and experienced teachers will choose to stay away from schools that are predominantly black in low-income areas. This also leads to a low average score. In addition, according to literature data, when the poverty variable is controlled, the ethnic factor will disappear. When the ethnic factor disappears, other related factors will also disappear.

After reading this analysis, the social status of women is still being suppressed, because from the data, it is known that the same low income headed by women, students in poor areas tend to have lower academic performance, which makes me think So, is it because the relative weakness of women leads to gang entanglement? Another important point is the issue of racial discrimination. What needs to be examined is why experienced teachers tend to avoid schools that are predominantly black in low-income areas. This is an invisible racial discrimination and an important factor

**3.Methodology**

3.1 Linear Regression Model

Linear regression is the most well-known modeling technique and one of the first choices when people learn how to predict models. In this technique, the dependent variable is continuous, and the independent variable can be continuous or discrete. The nature of regression is linear.

Linear regression establishes the relationship between the dependent variable (Y) and one or more independent variables (X) by using the best fitting straight line. The expression of linear regression model is below:

*Y = β0+ β1\*X1 …. + e*

（Y is dependent variable, X is independent variable, β0 is intercept of this regression, β1 is coefficient between Y and X, e is residual value.）

By establishing a linear regression model, we can generally see the impact of the factors we are studying on student performance. The next thing we must solve is how to evaluate the model we have established. We also need to notice some important condition below:

1. The independent variable and the dependent variable must satisfy a linear relationship.

2. Multiple regression has multiple collinearities, autocorrelation and heteroscedasticity.

3. Linear regression is very sensitive to outliers. Outliers will seriously affect the regression line and the final predicted value.

4. Multicollinearity will increase the variance of the coefficient estimates and make the estimates very sensitive to small changes in the model. The result is that the coefficient estimates are unstable.

5. In the case of multiple independent variables, we can use positive selection, backward elimination and stepwise selection methods to select the most important independent variable.

3.2 Least Square Method

Least squares method is a commonly used algorithm for fitting regression lines. It calculates the best fit straight line by minimizing the sum of squares of the vertical error between each data point and the predicted straight line. Because the calculation is the sum of squared errors, there is no mutual cancellation between the positive and negative errors. In the process of using Stata, we use the indicator R-square to evaluate the performance of the model. There has a question, why we choose Least Square Method? First, it is simple. It provides a measure of similarity with very good properties-it is non-negative; uniquely certain; it is symmetrical and has relevance to statistics and estimation theories.

In this project, the least square method (MSE) is more advantageous than the least absolute method (MAE). MAE can reduce the influence of discrete values or noise on the model, but we have made the data present a normal distribution after data cleaning, and will not produce discrete values, and the optimal solution of MSE is unique, and the project will not fall into Stagnant. In this research, we will use R-square as an indicator to compare the accuracy of each model.

3.3 Limitation of model

The disadvantage of linear model is that it cannot fit non-linear data well. So, we need to judge whether the relationship between variables is linear. At the same time, it can only simply predict the influence of the independent variable on the dependent variable and cannot accurately fit the regression as after training the model. Underfit often occurs. The data may be applicable to the polynomial regression model, but after using the linear regression model, it can only show a linear relationship. And using Stata for linear regression does not self-select the training set and test set, which will bias the predicted data, because the data size of machine learning and the data size of comparison are different.

3.4 Fixed Effect Model

Because we only want to compare the differences between specific categories or categories of each independent variable and the interaction effects between specific categories or categories of each independent variable, and we don’t want to infer that the same independent variable is not included in the other categories or categories within the experimental design, so we will use this model.

The first thing to consider is what the omitted variable is. Here we have a very clear result, race and region. We don’t want to get regression results with changes in race or region. What we want is to exclude race or region. With these two variables, the final regression result is what we want.

**4.Data Collection & Mining**

4.1 Collection

This data set is support by National Center for Education Statistics. Below is the description of the Dataset\par The Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K) focuses on children's early school experiences beginning with kindergarten and following children through middle school. The ECLS-K data provide descriptive information on children's status at entry to school, their transition into school, and their progression through 8th grade. The longitudinal nature of the ECLS-K data enables researchers to study how a wide range of family, school, community, and individual factors are associated with school performance. The children in ECLS-K came from both public and private schools and attended both full-day and part-day kindergarten programs. They came from diverse socioeconomic and racial/ethnic backgrounds. Also participating in the study were the children's parents, teachers, and schools.

4.2 Mining & Cleaning

In the data set given by the nation center for education statistics, there are some variables that will not affect the research results. Table 1.1 is statistic summary of data se. Our observations are 8105, there have no missing value in the data set. At the same time, we found that some data belongs to dummy variable. Some variables could be omitted. The invalid variables are removed through the operation on Stata to make the data set concise, and the data set has no missing values through Stata. This relevant data set mainly includes gender, race, region, parent marital status, student problems, extracurricular activities, time spent watching TV, and test score variables.

Table 1.2 are categorizing the variables that may be influencing factors and create a dummy variable to indicate whether the student participated in the event. We created a total of 4 dummy variables-whether there are extracurricular activities, whether there are social events, whether to watch TV. The linear regression shown by these variables can easily show their influence on students' academic performance.

Through regression statistics, we found that in addition to TV activities, other factors are statistically significant. So, in the next analysis, we will exclude the TV activity factor in the model, because it will not affect our model. At the same time, other variables are statistically significant. What we need to study is which part will affect students' academic performance. We could get an approach expression for this regression model:

*Total\_Grade = 280.3 + 15.27 \* extra\_activity + -7.992 \* problem + 11.83\*mom\_cur\_married+ other effect*

Other effect including but not limited to race, region, gender. These are the influencing factors that our regression model needs to consider, but our research is limited to other aspects, so we will group the data to eliminate the influence of this factor.

**5.Result**

5.1 Regression Result

In the table 1.3 The regression (1) results show that the mother’s marital status and the student’s grades have statistic significant. If student’s mom is current married, the total grade of student will increase by 14.81. The formula should be:

*Total Grade = 288.6 + 14.81 \* mom\_curr\_married + e*

(*Total grades* represent the sum of scores on math, reading and science. *e* represent other effect could be affected the total grade.)

The regression (2) results show that the regression on total grade and if student participate activity after lesson. We find out except art and dance activity, others is statistic significant, which means it has relationship with total grade. We also can get a formula for this regression; however, it may be inaccurate.

*Total Grade = 287.4 + -0.577\*part\_dance + 10.41\*part\_athletics + 6.467\*part\_club + 11.68\*part\_music + 0.503\*part\_art*

The regression (3) results show that the regression on total grade and if principal reports a problem. Excepting the problem of crime, all other problem is statistic significant. Let us make a conjecture, whether it is because the crime is a one-time event rather than a long-term event, so the results will not be impressed. It may be because of the occurrence of emotional crimes. In fact, the students' usual performance has been very good.

*Total Grade = 306.4 + -2.053\*problem\_crowding + -5.593\* problem\_turnover + -5.563\* problem\_parents + -4.855\* problem\_drugs + -9.225\* problem\_gangs + -1.221\* problem\_crime +-4.067\* problem\_weapons + -5.588\*parblem\_attacks*

After run regression on total\_grade and other variables, we figure out if student participates in art and dance lessons after school won’t affect their test score, on the other hand, it is not statistic significant. if the principal reports a problem with crime near campus and if the principal reports a problem with overcrowding also not statistic significant. The last regression model should be :

*Total Grade = 283 +3.593\*school\_type + 7.23\*part\_athletics + 5.195\*part\_club + 9.299\*part\_music + -4.510\* problem\_turnover + -2.904\* problem\_parents + - 4.872\* problem\_drugs + -7.460\* problem\_gangs + -2.473\* problem\_weapons + -3.083\*parblem\_attacks*

After we finish these regression models, we need to think carefully about whether there are any overlooked variables in the data we include. It is true that the region, gender, and ethnicity variables are ignored by us, so we need to filter the data and group the data so that the data is not biased.

In the next regression model, we need to think about the fixed effects of region and race. Set up a question, if the region changes and the race does not change, will there be changes in the performance of students with the same conditions? This is the essence of Fixed effects. Race and region are omitted variables, and we need to evaluate their impact on our regression model. By setting race and region as panel variables, and then using the Fixed-effects regression model, we can get the following regression results.

In the table 1.4 The regression (1) shows that the race fixed effect model, the regression (2) shows that the region fixed effect model. The sigma\_u and sigma\_e for regression (1) are 9.108 and 23.874. The sigma\_u and sigma\_e for regression (2) are 1.964 and 24.69. Because we have eliminated the effects of variables that we don’t want to see, this is the final regression result we want to get. When we fix the effect, the R-square becomes smaller, which shows that our model is more able to reflect What variables will affect the total grade variable.

**6.Conclusion**

The determinants of test scores for 5th grade students is divorce. If the mother is in a divorced state, the student’s grades will decline. At the same time, students who participate in extracurricular activities will have better grades than those who don’t. In addition, the poor campus environment will also reduce the student’s academic performance. Private school students It also tends to have better performance. Another phenomenon is shocking. Participating in music activities also has a great impact on students' performance, second only to the impact of parents' divorce on students. Even after we control the influence of regions and races. School crime incidents often fail to affect students. This makes us must pay attention to the mental health of students. Whether the students are accustomed to crimes or because they are indifferent to the criminals.

Reference

*Martin, B., & Wilson, C. (2000). Race, Poverty, and Test Scores: A Model of the Determinants of Test Scores in Toledo. The Negro Educational Review, 51(1-2), 23–36.*

*National Center for Education Statistics. (1999). Kindergarten Class of 1998-99. The data set of 5th students.*

**Table 1.1**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Obs | Mean | Std. Dev. | Min | Max |
| region | 8105 | 2.557 | 1.038 | 1 | 4 |
| gender | 8105 | 1.492 | .5 | 1 | 2 |
| race | 8105 | 1.764 | 1.08 | 1 | 4 |
| mom married at birth | 8105 | .784 | .411 | 0 | 1 |
| family income | 8105 | 58425.909 | 58918.039 | 1 | 999999.99 |
| part dance | 8105 | .138 | .345 | 0 | 1 |
| part athletics | 8105 | .675 | .469 | 0 | 1 |
| part club | 8105 | .303 | .459 | 0 | 1 |
| part music | 8105 | .314 | .464 | 0 | 1 |
| part art | 8105 | .112 | .315 | 0 | 1 |
| tv afternoon mf | 8105 | .668 | .81 | 0 | 6 |
| tv afterdinner mf | 8105 | 1.026 | .834 | 0 | 6 |
| tv saturday | 8105 | 2.808 | 1.711 | 0 | 23 |
| tv sunday | 8105 | 2.495 | 1.7 | 0 | 20 |
| problem crowding | 8105 | .192 | .394 | 0 | 1 |
| problem turnover | 8105 | .052 | .222 | 0 | 1 |
| problem parents | 8105 | .094 | .292 | 0 | 1 |
| problem drugs | 8105 | .215 | .411 | 0 | 1 |
| problem gangs | 8105 | .187 | .39 | 0 | 1 |
| problem crime | 8105 | .277 | .447 | 0 | 1 |
| problem weapons | 8105 | .132 | .339 | 0 | 1 |
| problem attacks | 8105 | .285 | .451 | 0 | 1 |
| reading test | 8105 | 100 | 10 | 63.087 | 117.653 |
| math test | 8105 | 100 | 10 | 66.521 | 117.086 |
| science test | 8105 | 100 | 10 | 69.526 | 120.411 |
| mom curr married | 8105 | .771 | .42 | 0 | 1 |
| family type | 8105 | 1.509 | .925 | 1 | 5 |
| (Table 1.1 is a statistical data analysis of the data set, please refer to the specific description from *Nation Center for Education Statistics*. ) | | | | | |

Table 1.2

|  |  |
| --- | --- |
|  |  |
| VARIABLES | total\_grade |
|  |  |
| extra\_activity | 15.27\*\*\* |
|  | (0.839) |
| problem | -7.992\*\*\* |
|  | (0.575) |
| tv\_activity | 2.870 |
|  | (2.738) |
| mom\_curr\_married | 11.83\*\*\* |
|  | (0.735) |
| Constant | 280.3\*\*\* |
|  | (2.864) |
|  |  |
| Observations | 8,105 |
| R-squared | 0.121 |

Robust standard errors in parentheses

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

Table 1.3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | (1) | (2) | (3) | (4) |
| VARIABLES | total\_grade | total\_grade | total\_grade | total\_grade |
|  |  |  |  |  |
| problem\_crowding |  |  | -2.053\*\*\* |  |
|  |  |  | (0.777) |  |
| problem\_turnover |  |  | -5.593\*\*\* | -4.510\*\*\* |
|  |  |  | (1.352) | (1.311) |
| problem\_parents |  |  | -5.563\*\*\* | -2.904\*\*\* |
|  |  |  | (1.083) | (1.034) |
| problem\_drugs |  |  | -4.885\*\*\* | -4.872\*\*\* |
|  |  |  | (0.979) | (0.848) |
| problem\_gangs |  |  | -9.225\*\*\* | -7.460\*\*\* |
|  |  |  | (1.015) | (0.911) |
| problem\_crime |  |  | -1.221 |  |
|  |  |  | (0.930) |  |
| problem\_weapons |  |  | -4.067\*\*\* | -2.473\*\*\* |
|  |  |  | (0.945) | (0.907) |
| problem\_attacks |  |  | -5.588\*\*\* | -3.083\*\*\* |
|  |  |  | (0.702) | (0.679) |
| part\_dance |  | -0.577 |  |  |
|  |  | (0.820) |  |  |
| part\_athletics |  | 10.41\*\*\* |  | 7.230\*\*\* |
|  |  | (0.654) |  | (0.640) |
| part\_club |  | 6.467\*\*\* |  | 5.195\*\*\* |
|  |  | (0.625) |  | (0.597) |
| part\_music |  | 11.68\*\*\* |  | 9.299\*\*\* |
|  |  | (0.606) |  | (0.582) |
| part\_art |  | 0.503 |  |  |
|  |  | (0.939) |  |  |
| mom\_curr\_married | 14.81\*\*\* |  |  | 9.401\*\*\* |
|  | (0.743) |  |  | (0.725) |
| school\_type |  |  |  | 3.593\*\*\* |
|  |  |  |  | (0.457) |
| Constant | 288.6\*\*\* | 287.4\*\*\* | 306.4\*\*\* | 283.0\*\*\* |
|  | (0.666) | (0.602) | (0.388) | (1.019) |
|  |  |  |  |  |
| Observations | 8,105 | 8,105 | 8,105 | 8,105 |
| R-squared | 0.052 | 0.103 | 0.083 | 0.182 |

Robust standard errors in parentheses

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

Table l.4

|  |  |  |
| --- | --- | --- |
|  | (1) | (2) |
| VARIABLES | total\_grade | total\_grade |
|  |  |  |
| mom\_curr\_married | 6.405\*\* | 9.419\*\* |
|  | (1.260) | (2.025) |
| part\_athletics | 5.397\*\*\* | 6.969\*\* |
|  | (0.458) | (1.332) |
| part\_club | 3.757\*\* | 5.017\*\* |
|  | (1.034) | (1.118) |
| part\_music | 8.712\*\*\* | 9.317\*\*\* |
|  | (0.881) | (0.345) |
| problem\_turnover | -1.657 | -3.966\*\* |
|  | (1.110) | (0.972) |
| problem\_parents | -1.873 | -3.168 |
|  | (2.136) | (1.536) |
| problem\_drugs | -5.297\*\* | -5.280\* |
|  | (1.163) | (2.220) |
| problem\_gangs | -3.392 | -6.504 |
|  | (2.362) | (3.564) |
| problem\_weapons | -1.484\* | -2.281 |
|  | (0.557) | (1.604) |
| problem\_attacks | -2.580\*\*\* | -3.222\*\* |
|  | (0.412) | (0.775) |
| school\_type | 2.814\*\* | 3.499\*\* |
|  | (0.533) | (1.007) |
| Constant | 286.9\*\*\* | 283.2\*\*\* |
|  | (0.503) | (0.659) |
|  |  |  |
| Observations | 8,105 | 8,105 |
| R-squared | 0.107 | 0.174 |
| Number of race | 4 |  |
| Number of region |  | 4 |

Robust standard errors in parentheses

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

(Table 1.4 is regression result of fixed effect model, group by race and region)