Asuncion Los Banos

DS-670 Capstone: Big Data & Business Analytics

Assignment 15

**Capstone of Asuncion Los Banos**

Good evening, my name is Asuncion Los Banos and tonight I will present my capstone, “Understanding the Influential Factors to Achievement Test Scores for Students in New Jersey”.

**Contribution of Competitor’s Article**

The title of my competitor’s paper is “The impact of family income on child achievement: Evidence from the earned income tax credit”. The authors, Dahl and Lochner, used a strategy of instrumental variable, or in other words they used a statistical method of correlation. They estimated the causal effect of income on children’s math and reading achievement. Their identification derived from large, nonlinear changes in the Earned Income Tax Credit (EITC) and data from Children of the NLSY which can be found from the Bureau of Labor Statistics. The National Longitudinal Surveys (NLS) are a set of surveys designed to gather information at multiple points in time on the labor market activities and other significant life events of several groups of men and women. For more than 4 decades, NLS data have served as an important tool for economists, sociologists, and other researchers.

**Description of Your Contribution**

My contribution for this capstone is to “Understand the factors that can influence a student’s test score in an achievement test.” I was in the search of answering: Do students who score high on standardized tests naturally good test takers or are there other factors that take place? If you went to school in the United States you will most have likely taken two major kinds of standardized tests: aptitude and achievements tests. For college bound high school students they will partake in an aptitude test that will predict how well they are likely to perform in educational settings. Those aptitude tests are the SAT and ACT, which forecast how well a high school student, will perform in college or university. Students while in elementary, middle, and high school will be exposed to the other type of standardized test, the achievement test. These score results are what citizens and school board members rely on to base on how well the school is performing.

**Data Source and Content**

The years for this capstone will be 2014, 2010, and 2007. Since I am searching for an influential factor on average test scores I decided to pick 2007 and 2010 because it was the start and end of the housing bubble and 2014 to view if there were any changes from the end of the bubble and four years later. The test scores will be from the 2014, 2010, and 2007 HSPA reading and math results. The HSPA was previously used to determine student achievement in reading, and mathematics as specified in the New Jersey Core Curriculum Content Standards. The data for test scores were retrieved from the State of New Jersey’s Department of Education website and these results will be my dependent variables. The dependent variables will be the test scores from the 21 counties in New Jersey. The HSPA scores for reading and math had a range of 100 to 300. Where a score of 199 and below indicated partially proficient, 200 to 249 indicated proficient and 250 to 300 indicated advanced proficient. For the purpose of this capstone the average score from the two subject areas will be used.

The independent variables of household income and ethnicity will be gathered from the U.S. Census Bureau’s American FactFinder. American FactFinder provides access to data about the United States and come from several censuses and surveys. I decided to utilize the Guided Search in the website to help search for the data I required. The constant filter for my search was the geographic type: County and New Jersey as the main state and all counties selected.

The household incomes are divided into four parts: households, families, married couple, and non family households. A household includes all the people who occupy a housing unit as their usual place of residence. A family household includes a householder and one or more people living in the same household who are related to the householder by birth, married or adoption. All people in a household who are related to the householder are regarded as members of his or her family. A family household may contain people not related to the householder, but those people are not included as part of the householder’s family. Therefore, the number of family households is equal to the number of families, but family households may include more members than do families. Not all households contain families since a household may compromise a group of unrelated people or people living alone. A married couple, as defined by the census, is a husband and wife enumerated as members of the same household. The married couple may or may not have children living with them. The number of married couples equals the count of married couple families related and unrelated married couple subfamilies. A non family household may contain only one person, the householder, or additional persons who are not relatives of the householder. Non family households may be classified as either female non family or male non family households.

The race used for this capstone is: African American, Asian, Caucasian, and Hispanic or Latino. When information was gathered from the American FactFinder the subject of Hispanic or Latino and Race was used because individuals can either be Hispanic or Latino or the either types of race. Besides the race mentioned above, there were other race that were not included in this capstone and they were: American Indian and Alaska Native, Native Hawaiian and Other Pacific Islander, some other race, and two or more races. These races were not added in the regression because they reflect very little of the population in New Jersey. Black or African American is a person having origins in any of the Black racial groups of Africa. It includes people who indicate their race as "Black, African American” or report entries such as African American, Kenyan, Nigerian, or Haitian. Asian is a person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent, including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam. It includes people who indicate their race as "Asian Indian," "Chinese," "Filipino," "Korean," "Japanese," "Vietnamese," and "Other Asian" or provide other detailed Asian responses. A person selecting Caucasian are: person having origins in any of the original peoples of Europe, the Middle East, or North Africa. It includes people who indicate their race as "White" or report entries such as Irish, German, Italian, Lebanese, Arab, Moroccan, or Caucasian. People who identify with the terms "Hispanic," "Latino," or "Spanish" are those who classify themselves in one of the specific Hispanic, Latino, or Spanish categories listed on the questionnaire as well as those who indicate that they are "another Hispanic, Latino, or Spanish origin."

**Your Method**

To model my data, I will be using linear regression as my statistical method. Linear regression can “predict” the value of the dependent variable based upon the values of one or more independent variables. This statistical data analysis will be used to determine where there is a linear relationship between a dependent variable and one or more independent variables. I will be utilizing two types of linear regression: simple and multiple linear regressions. Regression analysis has three major uses and they are: causal analysis, forecasting an effect, and trend forecasting. For my causal analysis I am trying to answer the following question: “What is the strength of relationship between average test score and average income?” Using regression to forecast an effect I am trying to answer: “How much more income does a household need to earn for a student to score higher in their test score?”

In a simple linear regression a single independent variable is used to predict the value of a dependent variable. In a multiple linear regression two or more independent variables are used to predict the value of a dependent variable. The only difference between a single linear regression and a multiple linear regression is the number of independent variables. However, for both there is only one dependent variable. My dependent variable of average test scores is measured in a continuous measurement scale of 0 to 300. The independent variables are also measured in a continuous measurement scale.

A single linear regression for each year will be completed for the different types of households and race. After completing a single regression on each type a multiple linear regression will be done with all types together. I have kept my independent variables to four variables to keep it from overfitting which can make the model inefficient. I am going to keep my model as simple as possible because statistically if the model includes a large number of variables the probability increases that the variables will be statistically significant from random effects.

The linear regression was completed using the R Interpreter for Apache Zeppelin. The formula for single linear regression for New Jersey counties average HSPA scores with the average income by household type:

avg\_combined\_scores = β0 + β1(Average income by household type)

The formula for multiple linear regression for New Jersey counties average HSPA scores with all average income by household type:

avg\_combined\_scores = β0 + β1(Household)+ β2(Families)

+ β3(Married Couple)+ β4(Non Family)

The formula for single linear regression for New Jersey counties average HSPA scores with race by type:

avg\_combined\_scores = β0 + β1(Race by type)

The formula for multiple linear regression for New Jersey counties average HSPA scores with all race:

avg\_combined\_scores = β0 + β1(African American)+ β2(Asian)

+ β3(Caucasian)+ β4(Hispanic or Latino)

**Quantitative Results 1**

The results for multiple linear regression on all average income by household type was statistically significant at all levels for 2014, but the results for 2010 and 2007 varied. In 2010, only married couple was statistically significant, but the overall p-value was 5.928e-07. In 2007, family and married couple were statistically significant with an overall p-value of 4.908e-07. The results showed that average income of married couple had the most influential factor for all three years. When you look at the table married couple is statistically significant for all years. When you view the linear regression on just married couple you can see that having a married couple household greatly influences test score.

* The regression for average income by for married couple in 2014 had a coefficient of 2.332e-04, an Adjusted R-squared of 0.5665 and p-value of 4.987e-05.
* The regression for average income by for married couple in 2010 had a coefficient of 2.382e-04, an Adjusted R-squared of 0.5164 and p-value of 0.0001464.
* The regression for average income by for married couple in 2007 had a coefficient of 2.088e-04, an Adjusted R-squared of 0.4445 and p-value of 0.0005775.
* The average income for married couples was all statistically significant for 2014, 2010, and 2007.

**Quantitative Results 2**

African American:

Before completing the regression for African American for 2014, 2010, and 2007 I hypothesized that test scores would be influential because there have been many articles regarding the “achievement gap”. After the single linear regression was completed I was not shocked with the results.

* The regression for African American in 2014 had a coefficient of -52.692, an Adjusted R-squared of 0.3059 and p-value of 0.005477.
* The regression for African American in 2010 had a coefficient of -58.868, an Adjusted R-squared of 0.3869 and p-value of 0.001552.
* The regression for African American in 2010 had a coefficient of -58.146, an Adjusted R-squared of 0.4775 and p-value of 0.0003146.
* The results for African American was statistically significant at all levels.

Asian:

* The regression for Asian in 2014 had a coefficient of 36.155, an Adjusted R-squared of 0.0217 and p-value of 0.2443.
* The regression for Asian in 2010 had a coefficient of 33.990, an Adjusted R-squared of -0.002513 and p-value of 0.342.
* The regression for Asian in 2007 had a coefficient of 28.095, an Adjusted R-squared of -0.01947 and p-value of 0.4414.
* The results for Asian showed we cannot conclude that a significant difference exists because the p-value for all three years were over 0.05.

However, because a regression is looking for the most influential factor it can mean that because Asian was not statistically significant it can mean that Asian American students excel in school because of culture. A study published in the journal PNAS, stated “Asian and Asian American youth are harder working because of cultural beliefs that emphasize the strong connection between effort and achievement. Studies show that Asian and Asian American students tend to view cognitive abilities as qualities that can be developed through effort, whereas white Americans tend to view cognitive abilities as qualities that are inborn”.

Caucasian:

* The regression for Caucasian in 2014 had a coefficient of 21.339, an Adjusted R-squared of 0.2111 and p-value of 0.02082.
* The regression for Caucasian in 2010 had a coefficient of 26.157, an Adjusted R-squared of 0.3005 and p-value of 0.005933.
* The regression for Caucasian in 2007 had a coefficient of 25.890, an Adjusted R-squared of 0.3448 and p-value of 0.003038.
* The regression results for Caucasian were statistically significant.

Hispanic or Latino:

Besides African American students falling in the “achievement gap”, Hispanic or Latino student’s average scores are significantly lower than average scores for White and Asian students.

* The regression for Hispanic or Latino in 2014 had a coefficient of -38.266, an Adjusted R-squared of 0.2295 and p-value of 0.01623.
* The regression for Hispanic or Latino in 2010 had a coefficient of -42.092, an Adjusted R-squared of 0.2461 and p-value of 0.0129.
* The regression for Hispanic or Latino in 2007 had a coefficient of -35.775, an Adjusted R-squared of 0.1971 and p-value of 0.02514.

The results for multiple linear regression on all races were statistically significant at all levels for all years. However, 2014 had the highest p-value of 0.004226, whereas 2007 had a p-value of 0.0001667.

**Discussion: Comparison With Your Competitor**

My competitor’s research correlated family income with test scores. They supported their hypothesis, linking test scores with household income, by using data based on five repeated measures of cognitive test scores per child. Using data derived from the U.S. government’s Earned Income Tax Credit (EITC) records, they identified nonlinear changes that indicated upward shifts, or expansions, in family income over twelve years (1987-1999), a twenty-percent increase of around $2,100, between 1993 and 1997. They applied an “instrumental variable strategy” that established a causal relationship between measurable expansions in family income and corresponding increases in children’s math and reading scores. The conclusions reached by recent studies suggest that unobserved heterogeneity and endogenous income shocks are important concerns. Furthermore, they suggest that income effects may be greatest among economically disadvantaged families. Dahl and Lochner completed a correlation on their study to find a cause and effect in family income and test scores. My approach was a regression where I used independent variables to find influential factors to a student’s test score.

The influential factor for a student’s test score is important to understand. Although a student cannot decide which race they are born into, understanding the powerful factor of household income is an important question. Regardless of household type, each one had a prominent factor to test scores. The results for average income by household and race shows positive influential factor between the variables and the average scores. African American and Hispanic students fall in the “achievement gap” where socio-economic factors including income levels, educational attainment, employment rates, housing options, neighborhood crime rates, and resources available to schools are worse for African Americans and Hispanics, on average, than for Whites (Patterson, C. J., et. al., 1990). The “achievement gap” often lead to fewer opportunities for African American and Hispanic children to access a wide range of activities and experience an enriched educational environment

**Performance on Big Data: Time Measurements**

The big chunk of preparing this capstone was gathering the data from State of New Jersey’s website and American FactFinder. Since the data was not “cleaned” it took me about 2 hours to prepare the data to run the linear regression in Zeppelin. Once the data was prepared loading them in Zeppelin was very quick and each linear regression took less than 1 second to complete.

**Conclusion**

Average income and race has an effective factor on achievement test scores, yet it may not be the main influencing factor. Students living in poor households may have a worse home environment or other characteristics that cannot be observed. Therefore, more study will need to be done to understand the influencing factor for test scores. However, while completing this capstone it made me wonder: How can the information in this capstone help in real life? So I decided to compile all the information and created a data visualization.

**Open Tableau**

How many of you in this room plan on getting married? How many are planning on purchasing your home? How about children? How many of you plan on having children? Now my next question is: do you know where the good school districts are in the State of New Jersey?

So as you can see in Tableau, is a map of New Jersey. All the little dots are a public schools ranging from Pre-K to 12th grade. Each school below to a DFG, which means District Factor Group. The DFGs represent an approximate measure of a community’s relative socioeconomic status (SES). The classification system provides a useful tool for examining student achievement and comparing similarly-situated school districts in other analyses. The DFGs were calculated using the following six variables that are closely related to SES:

1) Percent of adults with no high school diploma

2) Percent of adults with some college education

3) Occupational status

4) Unemployment rate

5) Percent of individuals in poverty

6) Median family income.

Therefore, a school closer to the beginning of the alphabet can be viewed as “low ranking” and schools later in the alphabet are “high ranking”.

Let’s make pretend we have a newly married couple and they are looking for a house. They can either select a range in median income or search for high ranking schools. So, let’s search for “J” DFG and see where they are located. Based on this filter we see that these are the cities with a “J” DFG. When you hover on the school information you will find information such as: school test results, student to teacher ratio, graduation rate, economically disadvantage students, etc. You can view this information and gauge if you would want to live in this city/county.