Asuncion Los Banos

DS-670 Capstone: Big Data & Business Analytics

Assignment 5

I will be using the 2014 and 2007 HSPA reading and math results. The data for 2014 and 2007 HSPA results 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. In 2014 and 2007 the numerical scores for reading and math 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. My independent variables of household income, ethnicity, married households, and education were gathered from the U.S. Census Bureau’s American FactFinder and have been recognized as having an influential impact on standardized test scores.

My defined variables:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Description** | **Label** | **Variable** | **Measure** | **Type** |
| Student’s Average Test Score | testscore\_avg | Average test score of reading and math HSPA | Numerical value | Dependent |
| Average Household Income | average\_income | Average income in a county of New Jersey | Dollars | Independent |
| Ethnicity | ethnicity\_white | Percentage of population identified as “White” in New Jersey | Percentage value | Independent |
| Married Households | married\_household | Percentage of population living in a married household | Percentage value | Independent |
| Education | education | Percentage of population who have achieved a bachelor’s or higher | Percentage value | Independent |

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 linear regression and multiple linear regression. 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 of 0 to 100 percent and dollars.

My single linear regression (SLR) will be regressing the average test scores of reading and math with average household income in New Jersey. I will have three multiple linear regressions (MLR I, MLR II, and MLR III) that will regress average test score with the independent variables factors (household income, ethnicity, married households, and education). MLR (I) will be regressing average test scores with white ethnicity and married households in the state. MLR (2) will be regressing average test scores with white ethnicity and education. MLR (3) will be regressing average test scores with white ethnicity and married households, and education. 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.

**Regression Models:**

**The formula for Single Linear Regression**

|  |
| --- |
| testscore \_ avg = β 0 + β 1 (average \_ income) |

Regressing New Jersey counties average HSPA scores with the average household income.

**The formula for Multiple Linear Regression (1)**

|  |
| --- |
| testscore \_ avg =  β 0 + β 1 (average \_ income) + β 2 (ethnicity \_ white) + β 3 (married \_ household) |

Regressing New Jersey counties average HSPA scores on the average household income, controlling for the percentage of the population in the county that is white and married households.

**The formula for Multiple Linear Regression (2)**

|  |
| --- |
| testscore\_avg =  β 0 + β 1 (average \_ income) + β 2 (ethnicity \_ white) + β 3 (education) |

Regressing New Jersey counties average HSPA scores on the average household income, controlling for the percentage of the population in the county that is white and with an educational level of some college or more.

**The formula for Multiple Linear Regression (3)**

|  |
| --- |
| testscore \_ avg =  β 0 + β 1 (average \_ income) + β 2 (ethnicity \_ white) + β 3 (married \_ household) + β 4 (education) |

Regressing New Jersey counties average HSPA scores on the average household income, controlling for the percentage of the population in the county that is white, married households, and adults with an educational level of some college or more.

I will be utilizing the Gauss–Markov theorem which states that in a linear regression model in which the errors have expectation zero and are uncorrelated and have equal variances, the best linear unbiased estimator (BLUE) of the coefficients is given by the ordinary least squares (OLS) estimator, provided it exists.

**Assumption 1 - Linear in Parameters**

The model is in linear parameters, but does not have to be linear in the x’s: Y = β 0 + β 1 X1 + β i Xi

**Assumption 2 - Random Sample of n Observations**

An appropriately sized, random sample is used in the regression model.

**Assumption 3 – Zero Conditional Mean**

The mean of the error terms has an expected value of zero given values for the independent variables.

**Assumption 4 – No Perfect Collinearity**

The error term, *u*, is independently distributed and not correlated with any of the variables. The

variables are not correlated. The assumption of no perfect collinearity states that *there is no exact linear relationship among the independent variables*. This assumption implies two aspects of the data on the independent variables.

**Assumption 5 - Homoskedasticity**

The error terms all have the same variance and are not correlated with each other. In statistical jargon, the error terms are independent and identically distributed (iid). This assumption means the error terms associated with different observations are not related to each other.

**Data Collection:**

The data for test score will be retrieved from the State of New Jersey’s Department of Education website and gathering it from the assessment reports section. I am downloading the High School Proficiency Assessment Spring 2014 and High School Proficiency Assessment Spring 2007 data. Each school is located in one of the twenty-one county in New Jersey. The data files have columns for mean total reading and mean total math for each school. I will need to combine the reading and math score for each school and then gather the average the combine score by county. Some public schools are considered charter schools and are not linked to a county. I will need to find the county for these schools myself and add it in my data file.

Pseudo code to run in Zeppelin to combine the reading and test scores for 2014 and 2007:

% r

2014 \_ hspa < - read. csv ('2 014 hspa. csv ', header = TRUE, sep = ", ")

2014 \_ reading < - as. numeric (2014 \_ hspa $ col 1)

2014 \_ math < - as. numeric (2014 \_ hspa $ col 2)

reading \_ math \_ 2014 < - 2014 \_ reading + 2014 \_ math

% r

2007 \_ hspa < - read. csv ('2 007 hspa. csv ', header = TRUE, sep = ", ")

2007 \_ reading < - as. numeric (2007 \_ hspa $ col 1)

2007 \_ math < - as. numeric (2007 \_ hspa $ col 2)

reading \_ math \_ 2007 < - 2007 \_ reading + 2007 \_ math

The data for my independent variables will be retrieved 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, New Jersey as the main state and all counties selected. Each independent variable can be easily located on the website using the ID column.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Year** | **Label** | **Table, File or Document Title** | **Dataset** | **ID** |
| 2014 | average\_income | Income In the Past 12 Months (In 2015 Inflation - Adjusted Dollars) | 2011-2015 American Community Survey 5-Year Estimates | S1901 |
| 2014 | ethnicity \_ white | RACE  Universe: Total population | 2015 American Community Survey 1 - Year Estimates | B02001 |
| 2014 | married \_ household | Households and Families | 2011 - 2015 American Community Survey 5-Year Estimates | S1101 |
| 2014 | education | Educational Attainment | 2011 - 2015 American Community Survey 5 - Year Estimates | S1501 |
| 2007 | average \_ income | Income in the Past 12 Months (In 2007 Inflation - Adjusted Dollars) | 2005 - 2007 American Community Survey 3 - Year Estimates | S1901 |
| 2007 | ethnicity\_ white | RACE  Universe: Total population | 2007 American Community Survey 1 - Year Estimates | B02001 |
| 2014 | married \_ household | Households and Families | 2005 - 2007 American Community Survey 3 - Year Estimates | S1101 |
| 2007 | education | Educational Attainment | 2005 - 2007 American Community Survey 3 - Year Estimates | S1501 |

For average income the pseudo code to run in Zeppelin to gather the value for each county in 2014 and 2007:

% r

avgincome \_ 2014 <- read.csv('avgincome \_2014.csv', header=TRUE, sep = "," )

avgincome \_ per\_ 2014 <- avgincome\_2014$mean \_ income

For race I will divide White alone from Total population from the data set. Pseudo code to run in Zeppelin to gather the percentage value for ethnicity for each county in 2014 and 2007:

% r

ethnicity \_ 2014 <- read.csv('ethnicity \_2014.csv', header=TRUE, sep = "," )

ethnicity \_ per\_2014 <- ethnicity\_2014$white \_ alone / ethnicity \_ 2014 $ total \_ population

% r

ethnicity \_ 2007 <- read.csv('ethnicity \_2007.csv', header=TRUE, sep = "," )

ethnicity \_ per\_2007 <- ethnicity\_2007$white \_ alone / ethnicity \_ 2007 $ total \_ population

For race I will divide married household from total population from the data set. Pseudo code to run in Zeppelin to gather the percentage value for married households for each county in 2014 and 2007:

% r

married \_ 2014 <- read.csv('married \_2014.csv', header=TRUE, sep = "," )

married \_ per\_2014 <- ethnicity\_2014 $ married \_ estimate / ethnicity \_ 2014 $ total \_ households

% r

married \_ 2007 <- read.csv('married \_2007.csv', header=TRUE, sep = "," )

married \_ per\_2007 <- ethnicity\_2007 $ married \_ estimate / ethnicity \_ 2007 $ total \_ households

To gather the percentage value for education attainment I will need to combine Population 18 to 24 years - Bachelor's degree or higher + Population 25 years and over - Bachelor's degree Population 25 years and over - Graduate or professional degree. Pseudo code to run in Zeppelin to gather the percentage value for education for each county in 2014 and 2007:

% r

education \_ 2014 <- read.csv('education\_2014.csv', header=TRUE, sep = "," )

married \_ per\_2014 <- education \_ 2014 $ column \_ x + education \_ 2014$ column \_ y + education \_ 2014 $ column \_ z

% r

education \_ 2007 <- read.csv('education\_2007.csv', header=TRUE, sep = "," )

married \_ per\_2007 <- education \_ 2007 $ column \_ x + education \_ 2007$ column \_ y + education \_ 2007 $ column \_ z

I will need to combine all these data together to be able to run a linear regression. The pseudo code to run in Zeppelin to bind these data:

% r

independent\_variables\_2014 <- rbind ( avgincome \_ 2014, ethnicity \_ 2014, married \_ 2014, education \_ 2014)

% r

independent\_variables\_2007 <- rbind ( avgincome \_ 2007, ethnicity \_ 2007, married \_ 2007, education \_ 2007)