Asuncion Los Banos

DS-670 Capstone: Big Data & Business Analytics

Assignment 10

**Results**

A single linear regression was completed for all the different types of household average incomes for 2014, 2010, and 2007. Then a multiple linear regression was completed on all household average incomes and the results were very interesting.

Household:

* The regression for average income by household in 2014 had a coefficient of 3.277e-04, an Adjusted R-squared of 0.7362 and p-value of 4.017e-07.
* The regression for average income by household in 2010 had a coefficient of 3.405e-04, an Adjusted R-squared of 0.6904 and a p-value of 1.855e-06.
* The regression for average income by household in 2007 had a coefficient of 2.953e-04, an Adjusted R-squared of 0.6374 and a p-value of 8.727e-06.
* This means that the average income of household has an influential factor where if there was a $10,000 increase in average household income of the county there would be a 3.28, 3.41, and 2.95 increase in the average HSPA scores for 2014, 2010, and 2007 respectively.

|  |  |
| --- | --- |
| **2014** |  |
| **2010** |  |
| **2007** |  |

Table 1: Linear regression of average income by household

Family:

* The regression for average income by for a family household in 2014 had a coefficient of 2.672e-04, an Adjusted R-squared of 0.7341 and p-value of 4.336e-07.
* The regression for average income by for a family household in 2010 had a coefficient of 2.875e-04, an Adjusted R-squared of 0.6941 and p-value of 1.678e-06.
* The regression for average income by for a family household in 2007 had a coefficient of 2.494e-04, an Adjusted R-squared of 0.6318 and p-value of 1.014e-05.
* Again, a family household has an influential factor where if there was a $10,000 increase in average family income of the county there would be a 2.67, 2.87, and 2.49 increase in the average HSPA scores for 2014, 2010, and 2007 respectively.

|  |  |
| --- | --- |
| **2014** |  |
| **2010** |  |
| **2007** |  |

Table 2: Linear regression of average income by family

Married Couple:

* 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 were all statistically significant for 2014, 2010, and 2007.

|  |  |
| --- | --- |
| **2014** |  |
| **2010** |  |
| **2007** |  |

Table 3: Linear regression of average income by married couple

Non Family:

* The regression for average income by for a non family household in 2014 had a coefficient of 5.742e-04, an Adjusted R-squared of 0.5821 and p-value of 3.484e-05.
* The regression for average income by for a non family household in 2010 had a coefficient of 5.435e-04, an Adjusted R-squared of 0.423 and p-value of 0.0008449.
* The regression for average income by for a non family household in 2007 had a coefficient of 4.607e-04, an Adjusted R-squared of 0.3963 and p-value of 0.00133.
* The coefficient results for a non family household were very surprising for me because non family households contain one person, the householder and additional persons who are not relatives of the householder. Therefore, an income increase has higher influential factor for student who might be in this type of household because their test scores increases higher than the other types of household. Just think about what type of student is living in a non family household, a foster child.

|  |  |
| --- | --- |
| **2014** |  |
| **2010** |  |
| **2007** |  |

Table 4: Linear regression of average income by non family

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.

|  |  |
| --- | --- |
| **2014** |  |
| **2010** |  |
| **2007** |  |

Table 5: Multiple linear regression of average income by all household types

After completing the regression for average income by households, a regression for race was completed. The same method was done for race where a single a single linear regression was completed for the different types of race for 2014, 2010, and 2007 and then a multiple linear regression of all race was completed.

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.

|  |  |
| --- | --- |
| **2014** |  |
| **2010** |  |
| **2007** |  |

Table 6: Single linear regression for African American

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.

|  |  |
| --- | --- |
| **2014** |  |
| **2010** |  |
| **2007** |  |

Table 7: Single linear regression for Asian

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.

|  |  |
| --- | --- |
| **2014** |  |
| **2010** |  |
| **2007** |  |

Table 8: Single linear regression for Caucasian

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.

|  |  |
| --- | --- |
| **2014** |  |
| **2010** |  |
| **2007** |  |

Table 9: Single linear regression for Hispanic or Latino

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.

|  |  |
| --- | --- |
| **2014** |  |
| **2010** |  |
| **2007** |  |

Table 10: Multiple Linear Regression for all race