California Math Test Regression Analysis

Michael Cultrera

Colorado State University – Global Campus

California Math Test Regression Analysis

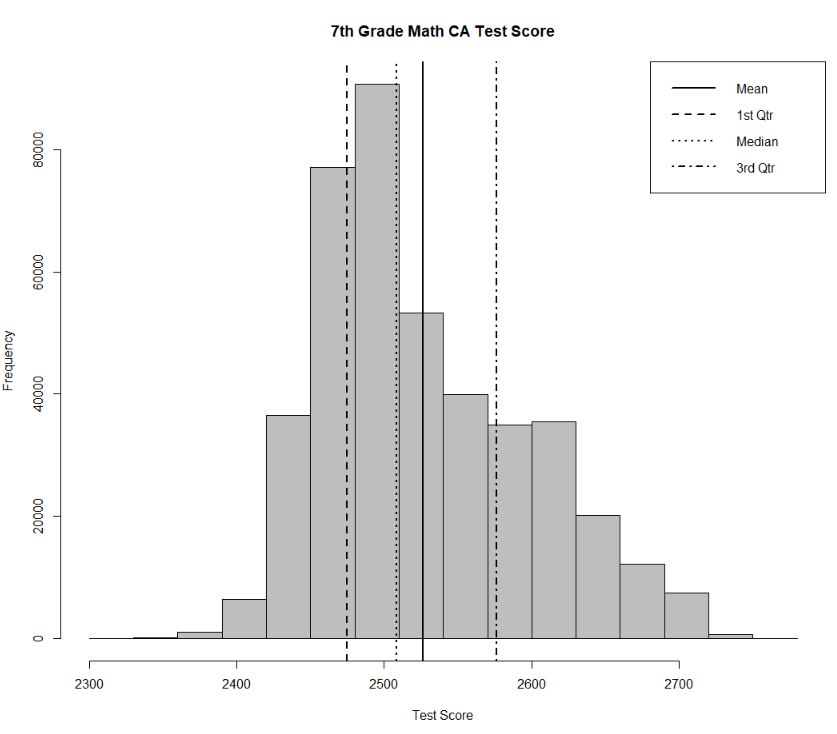
The state of California conducts annual testing for students in the public schools. The California Assessment of Student Performance and Progress, “CAASPP”, includes a language arts and mathematics assessment for grades three through eight and grade eleven. This analysis focuses on the math assessment done on seventh graders in the 2018-2019 school year. The state of California provides the test results data online based on pre-defined subgroups with scores removed when the subgroup population is too small to avoid the potential revelation of individual student information. In addition to testing data, the zip code for the school can be leveraged to pull in useful information regarding the surrounding areas. After identifying the most useful subgroup class defined by the testing dataset, a multi-variate regression analysis was done to incorporate many different factors based on school’s geography that is able to explain up to 88% of the variability in test scores. Student performance and capability to learn is essential in society for individual development. This analysis begins targeting factors that may be impacting test scores as well as suggesting different avenues for subsequent studies.

**California Test Data**

The California testing data is robust, but there are two key limitations that impacted the analysis. Due to privacy laws, FERPA in particular, any data that potentially could be used to reveal individual student information must be withheld. As a result, individual scores are not provided, but the mean of pre-defined subgroups are provided. The number of students reporting under the subgroups is included, so proper weighting between schools was maintained. However, the use of subgroup means, in lieu of individual scores, will naturally reduce score variance.

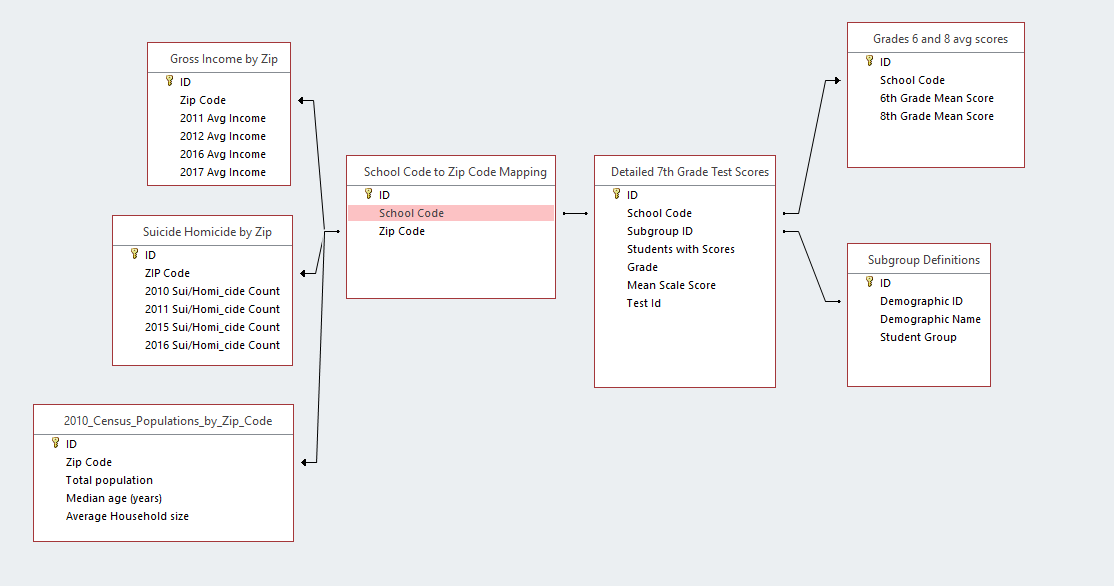
California collects and maintains many demographic attributes such as gender, ethnicity, economic disadvantage, disability, parent’s education level, and military status. However, these fields are all classified under a single subgroup field. There is a subgroup classification field that allows the data to be viewed using the different demographics such as gender and disability status. For example, it is possible to see the mean of scores of every student who identified as male or female separately. It is also possible to see the mean of scores of every student who identified as having or not having a disability. It is not possible to cross analyze these categories to see the performance of males who have a disability. Presumably, this is intentionally done to avoid incidentally revealing individual information through cross-sample analysis. The sole exception to the multiple demographic analysis is ethnicity joined with economic disadvantageousness. These subgroups were cross-analyzed and included in the published dataset.

Although the published data included county and school district identification, in order to keep the source data as detailed as possible, every record pertained to a single school and subgroup. The mean scale score is the dependent variable for this analysis. Figure 1 below illustrates the central tendency and distribution of the scores. The histogram illustrates the scores are represented with a right leaning skew and an overall mean score of 2,526. The median is 2,509. The standard deviation is 69.7, and the inner quartile range is 2,475 to 2,576.

**

*Figure 1. Histogram prepared in R Studio illustrates the distribution of test scores*

In addition, to the California testing dataset, there were three additional datasets joined based on the school’s zip code. The datasets included tax information to provide the average income, the suicide and homicide counts, population size data, median age, and average household size. Microsoft Access was leveraged to reduce the files to the required fields and joined together to form a single dataset. Figure 2 illustrates the table relationship mapping.

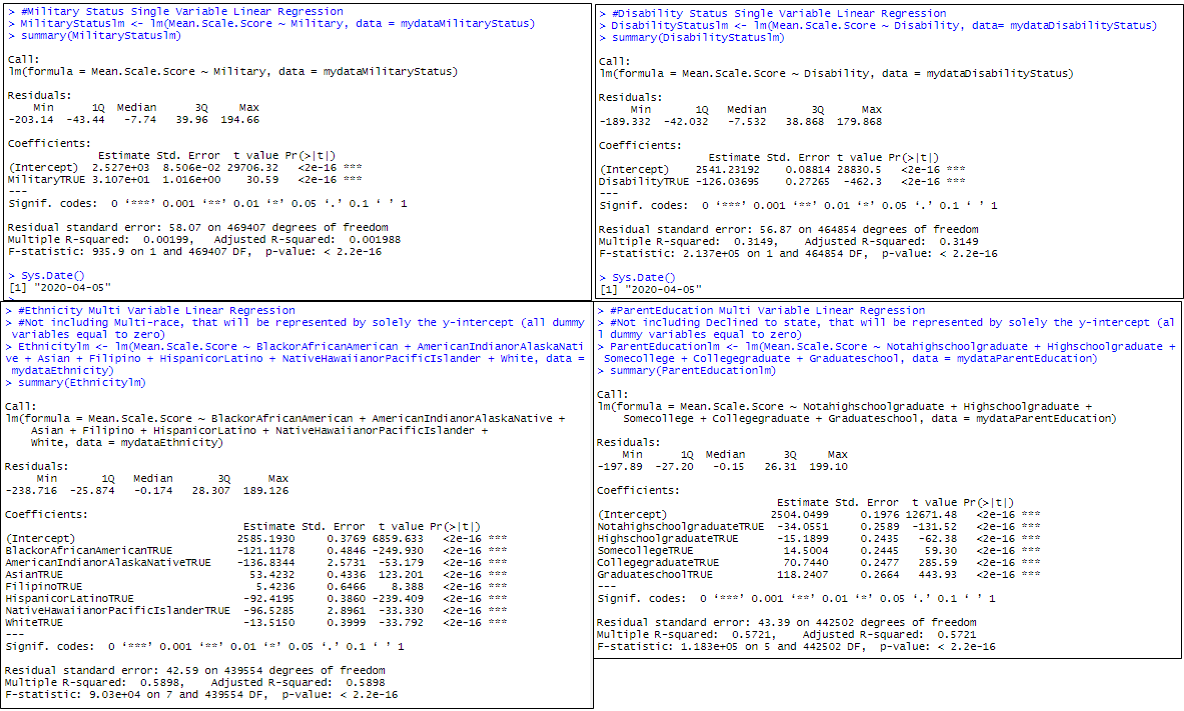
**

*Figure 2. Microsoft Access table relationships mapped to create a single dataset*

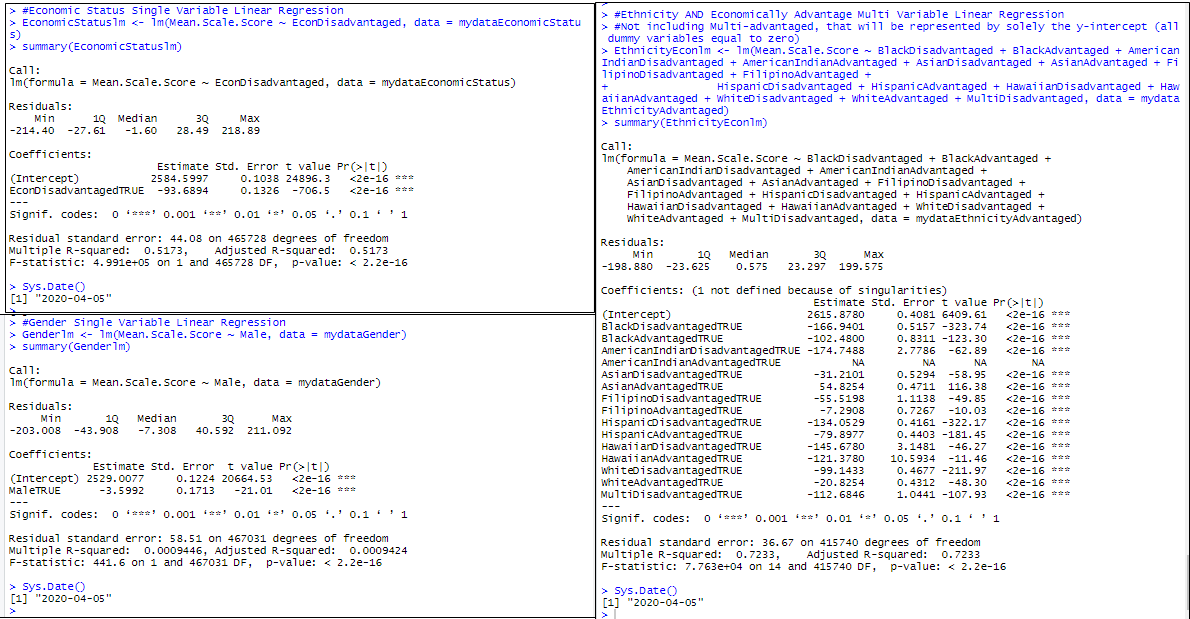
**Determining the Best Subgroup Classification**

As discussed previously, the California testing data does not permit the cross-analysis of different demographic subgroups with the notable exception of ethnicity and economic disadvantageousness. Since only one classification can be leveraged, multiple linear regression models were developed to identify the best one. Dummy variables were created in R Studio to allow for the categorical independent variables to be analyzed with linear regression. Figures 3 and 4 below outline the code and output from R Studio for the seven different regression models based on the seven different subgroup classifications.

All seven models reported p-values less than .05 suggesting there is a significant relationship, although the R2 values ranged from less than 1% to 72%. The model with the highest R2 value was the cross classification between ethnicity and economic advantageousness. The R2 value indicates that 72% of the variability in testing scores is explained by the ethnicity and economic subgroup. This subgroup classification will be leveraged exclusively in the subsequent analysis and unfortunately, any additional explanatory power that the other variables have will not be accessible.

**

*Figure 3. Represents the code in R Studio to generate the different linear regression models based on subgroup classifications. Starting in the top-left and moving clockwise: military status, disability status, ethnicity, and parent education.*



*Figure 4. Represents the code in R Studio to generate the different linear regression models based on subgroup classifications. Starting in the top-left and moving clockwise: economic disadvantageousness, ethnicity crossed with economic disadvantageousness, and gender.*

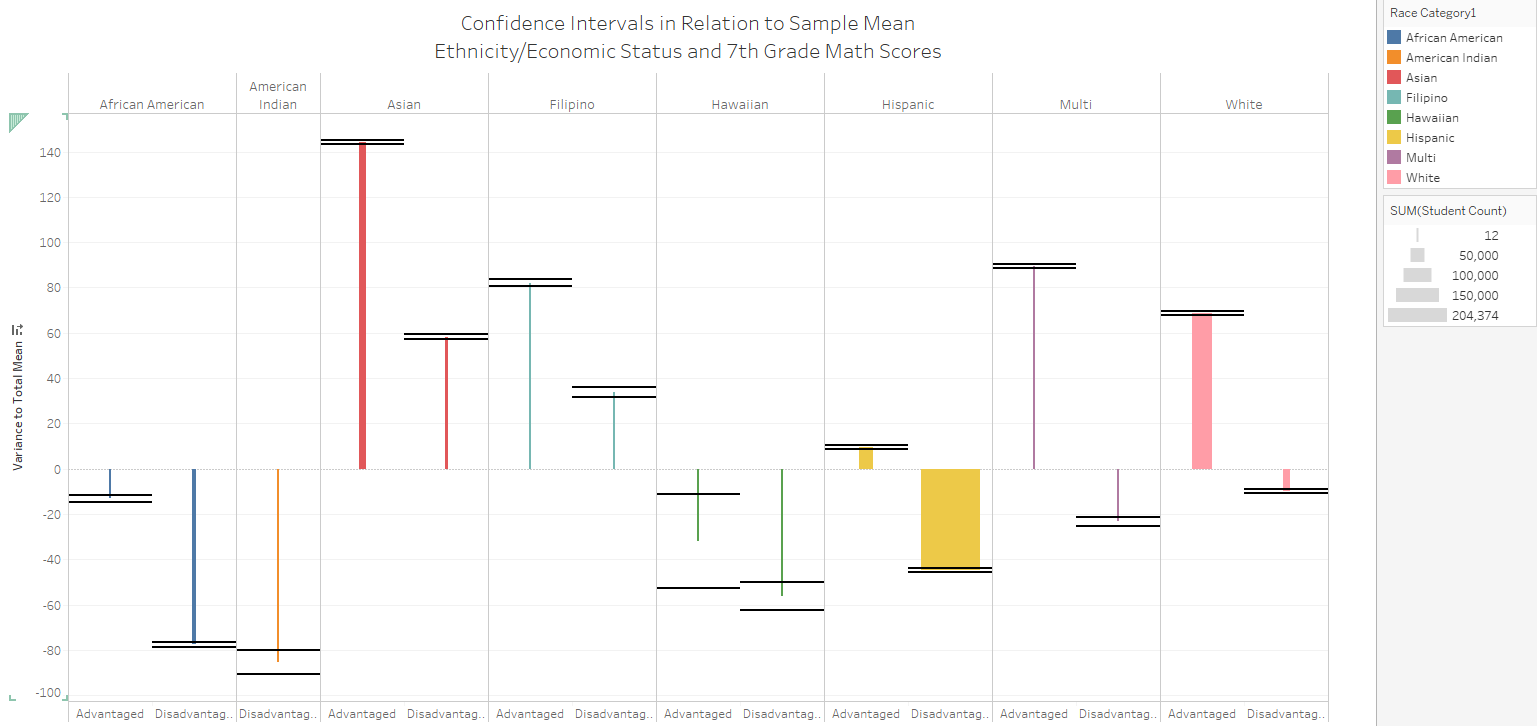
**Investigating the Ethnicity and Economic Disadvantageousness Classification**

Before enhancing the regression model with additional variables, it is worth reviewing the disparate impact that ethnicity and economic group has on the testing scores in relation to the total sample mean. Figure 5 below is a chart prepared in tableau and can be found here (along with all other Tableau charts shown in this analysis: (<https://public.tableau.com/views/PortfolioProjectEthnicityAdvantaged/Story1?:display_count=y&publish=yes&:origin=viz_share_link>). This visualizes the difference between the subgroups as well as how much of the data is represented within the subgroups by the width of the bar. The 95% confidence intervals are attached to the bars indicating the true population mean is within those bars. There are a few immediate takeaways. For example, in every ethnic category, the economically disadvantaged group performs worse. There are certain ethnic groups that still average below the total mean even when only considering those not economically disadvantaged such as African Americans, American Indians, and Native Hawaiians. The reverse is true for the Asian and Filipino ethnic groups in that even the economically disadvantaged are averaging above the total mean.

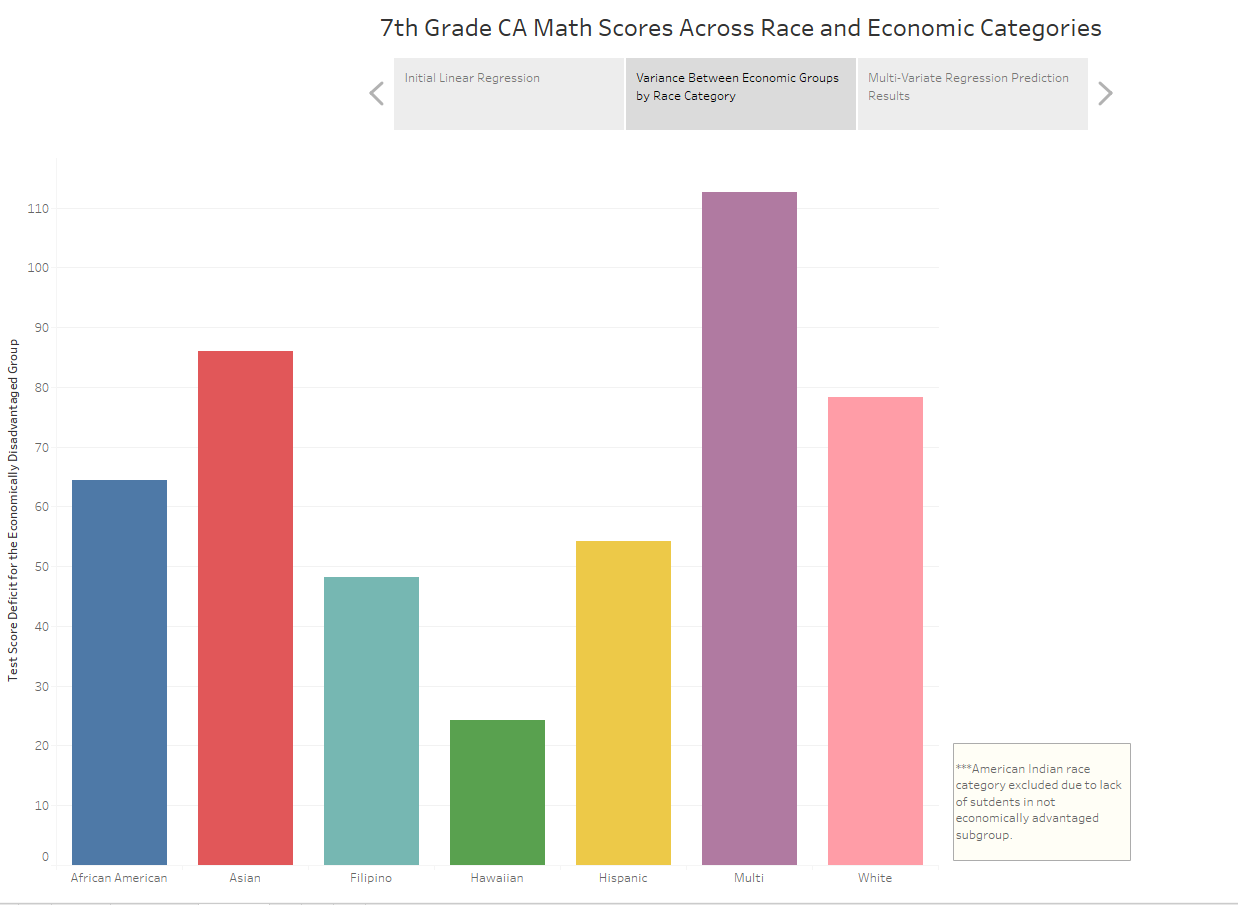
In addition to the variance from the mean, the amount of students within each subgroup is illustrated with width of the bars. The largest subgroup is Hispanic and economically disadvantaged followed by White and not economically disadvantaged. The less populated subgroups also have a wider range for their confidence intervals as would be expected.

The data represented within each category Although both ethnicity and economic group clearly have an impact, it appears that the dominant factor is ethnicity. This is confirmed by the individual regression models shown in Figures 3 and 4. The economic-only model generated an R2 value of 52%, but the ethnicity-only model generated an R2 value of 59% highlighting ethnicity as the dominant factor.

An interesting analysis that the cross-subgrouping allows for is the relative impact of economic status across the different ethnicity groups. Figure 6 is a chart prepared in Tableau that visualizes the differences. The standout ethnic groups are the Hawaiian and multi race categories. The former for its relatively small difference and the later for its large difference. Frequently in the media, ethnicity and economic status are viewed separately, but understanding how these demographics impact each other differently may be an area for future studies.

**

*Figure 5. Chart prepared in Tableau to visualize the ethnic and economic subgroups differences across test scores.*

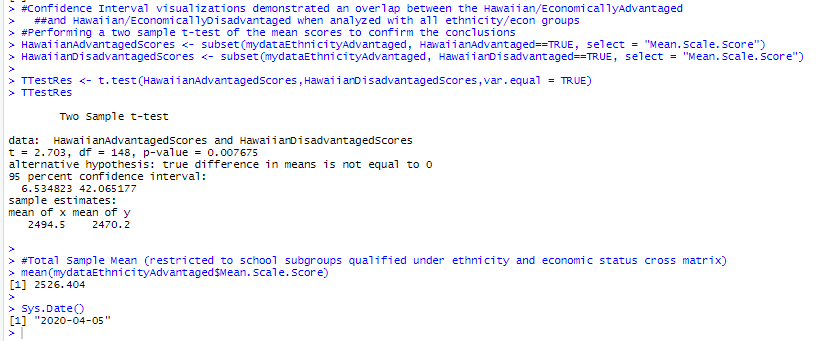
* Figure 6. Chart prepared in Tableau to visualize the difference between the two economic classes across the different ethnic categories.*

Revisiting Figure 5, an interesting aspect was the overlapping confidence intervals for Hawaiian disadvantaged and Hawaiian not disadvantaged. Overlapping confidence intervals suggest that the population mean is not statistically different with a 95% certainty. To further investigate this, a two sample t-test analysis was performed in R Studio. The code and output generated can be found in Figure 7 below. The hypotheses being tested are:

HO: The Hawaiian economically disadvantaged and Hawaiian not economically disadvantaged groups have equal population means for test scores.

HA: The Hawaiian economically disadvantaged and Hawaiian not economically disadvantaged groups have different population means for test scores.

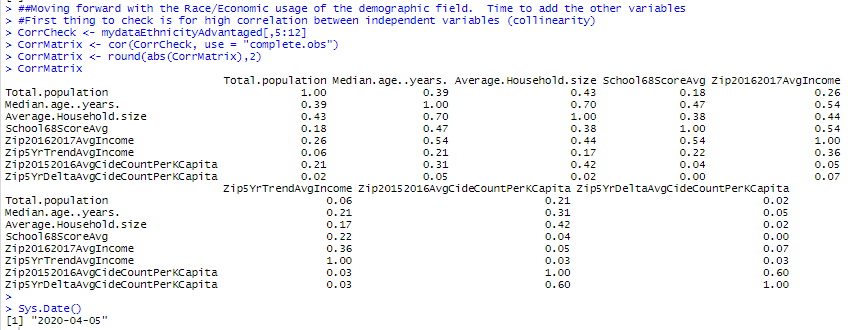
The results of the t-test shown in Figure 7 provide a p-value of .007 which is below the .05 figure necessary to suggest a significant difference in population means. Therefore, the null hypothesis is rejected and the subsequent models will continue to keep the Hawaiian ethnicity group separated by economic advantageousness.



*Figure 7. Represents the code in R Studio to generate the Two Sample T-Test*

**Incorporating Additional Data Points**

Now that the optimal subgroup classification is narrowed to ethnicity crossed with economic disadvantageousness, additional data points sourced from alternate datasets can be added to the multi-variate linear regression. Prior to inputting the additional variables, an important consideration is the risk for collinearity, that is the high correlation between independent variables. The presence of significant collinearity can lead to inaccurate model results. The sample size is pretty large with almost 416,000 records, which helps avoid collinearity. To investigate the potential for collinearity, a correlation matrix was produced as shown in Figure 8 below. The largest correlation found is between median age of household and average household size with a correlation coefficient of .7, which is low enough to leave both variables in the regression model.

 *Figure 8. Represents code and output from R Studio for the correlation matrix of the additional fields based on zip code.*

An important factor when considering test performance is the school’s overall performance. The grade being analyzed is the seventh grade. Commonly, junior high schools in America cover grades six through eight. Rather than average all three grades to develop a school performance variable, only grades six and eight were averaged to avoid incidental correlation.

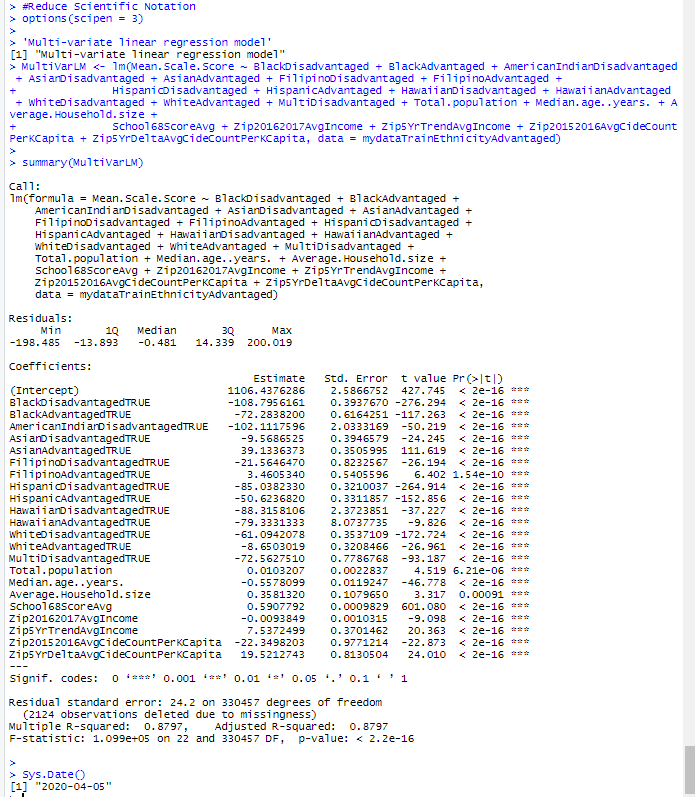
The economic disadvantageousness subgroup is significant and conveys financial information. However, the average income by zip code conveys financial information as a linear variable, whereas the subgroup is limited as a Boolean dummy variable. Additionally, the zip code income reveals information for the general population and not specifically for the student. A multi-variate regression including both variables can provide insight into the relationship between the variables and their relative impact on the test scores. To smooth out yearly fluctuations the income information is averaged between two years. The latest year available was 2016, so the current income is represented by an average between 2015 and 2016. Another potential factor is the income trend of a given area. If an area is becoming progressively poorer or richer overtime, that can certainly lead to school performance. To capture this concept, the 2010 and 2011 income was averaged and a percentage change over 5 years was calculated.

Another variable to consider is the crime and suicides within an area. It’s reasonable to consider that a high crime area may negatively impact nearby school performance. Similar to the income figures discussed above, two years were averaged to smooth out any single year outliers. The suicide and homicide counts were combined and a ratio was calculated by the population reported on the 2010 census for the zip code. The current state of homicides and suicides is a variable to be included in the model, but a delta between 2010/2011 and 2015/2016 is also considered to stand in for the trend of crime and suicides.

Lastly, three new variables were incorporated based on the 2010 census by zip code: total population, median age, and average household size. Each of these will be incorporated as their own independent variable. Future studies will be able to incorporate the 2020 census information and will be able create trend variables similar to those created for income and crime.

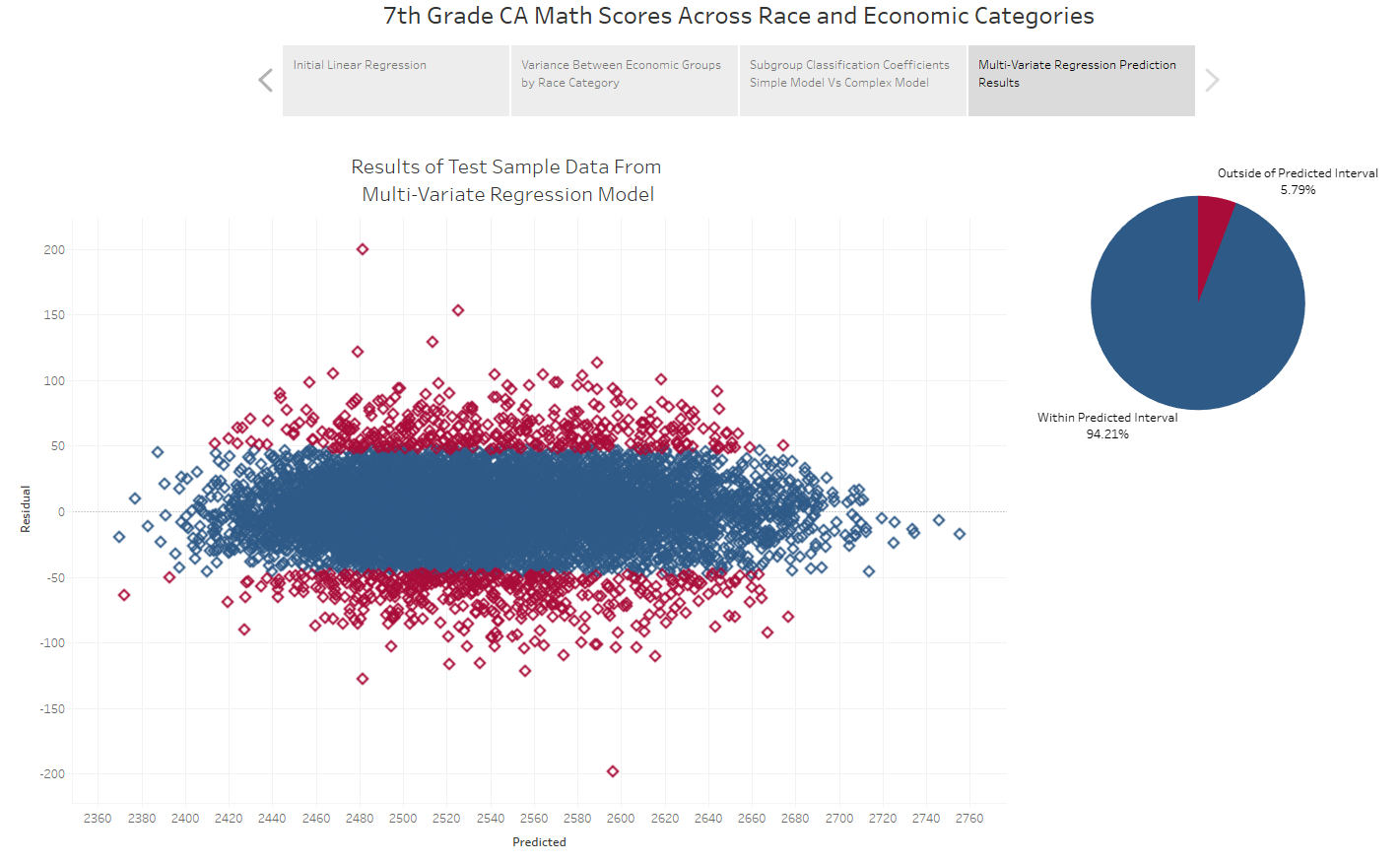
**Final Multi-Variate Regression Analysis**

Prior to the final regression model being produced the dataset was randomly split into a training data set and testing data set which represented eighty and twenty percent of the total records respectively. The final multi-variate regression model was produced in R Studio and can be seen in Figure 9. All independent variables were found to have a significant impact on the dependent variable test scores. The testing data set was then used to predict and evaluate over 82 thousand student’s test scores with an accurate prediction in 94.2% of tests. The code used to generate the predicted data set is shown in Figure 10. The predictive performance of the model on the testing data is visualized in Figure 11.

 *Figure 9. Represents code and output from R Studio for the final regression model.*



*Figure 10. Represents code from R Studio that applied the testing data set to the second model to predict and evaluate the model’s performance*

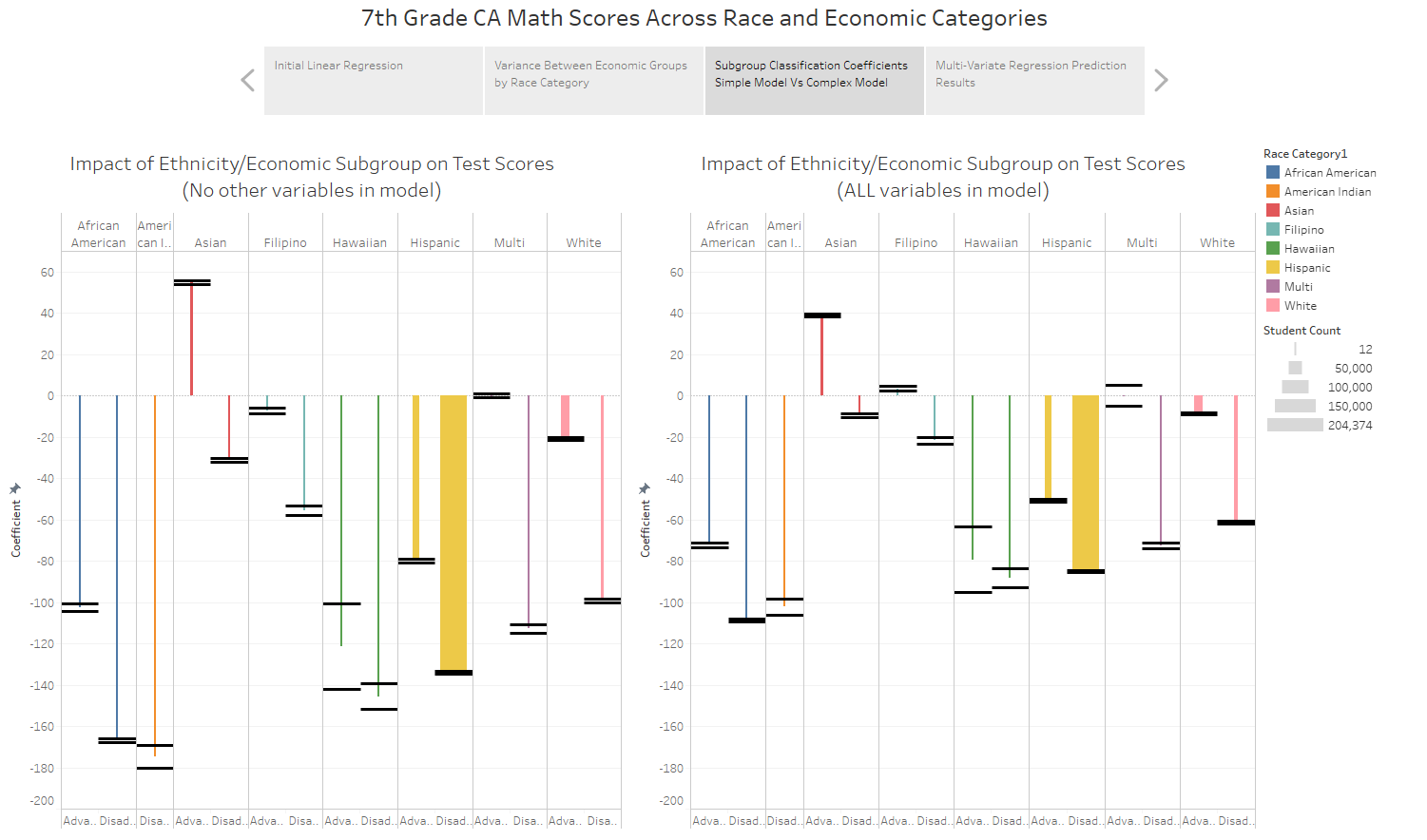
**

*Figure 11. Tableau chart illustrating the new model accurately predicted 94.2% of the testing data’s scores as measured by the prediction intervals.*

As is often the case with regression models, the y-intercept doesn’t provide much insight as a student registering zeroes for all independent variables is not realistic. Since the scale is very different between the independent variables, the coefficients of the model also don’t provide an instant comparison point.

The dummy variables for the ethnicity and economic subgroups do allow for some interesting insight. The larger the coefficient among the category variables, the larger impact that subgroup has on testing score. You can reliably draw ordinal conclusions from the coefficients. For example, the coefficient for the White disadvantaged group is -61.1 and the coefficient for Filipino disadvantaged group is -21. It is conclusive to say that belonging to the white disadvantaged group has a larger predicted negative impact on testing score than the Filipino disadvantaged. However, any magnitude comparisons are invalid. It would be inaccurate to conclude that the impact from the former group is approximately three times greater than the latter. The chart in Figure 5 from the first model was designed to illustrate the impact in reference to the sample mean. A similar chart cannot be produced from the final multi-variate model due to the additional variables.

There are important insights that can also be gathered by comparing the coefficients between the two models. The changes in the coefficients is visualized in Figure 12 below. The chart illustrates that for every subgroup the coefficients were reduced. This makes intuitive sense as the additional variables in the second model captured some of the impact that was presumed to be solely a result of the ethic and economic subgroup.



*Figure 12. Tableau chart to show the coefficients comparison between the first model where the subgroups were the only independent variables, and the second where other variables were considered.*

**Conclusion**

The California School testing program provides a good source of information to understand many variables’ impact on student performances in schools. The data had limitations such as reporting only subgroup mean scores rather than individual scores which likely reduced the true population variance. Additionally, there may be an important consideration for how being an extreme minority at a particular school impacts performance. Does being one of only a handful of an ethnic group at a particular school have a strong impact on performance? That is a worthwhile question to pursue, but unfortunately, with the publicly available data redacted in accordance with FERPA, there is no information available for these extreme minority cases as any inclusion of their scores could reveal their personal information.

The other key limitation to the data is the consolidation of demographics into a single subgroup field. Fortunately, the preparers of the data did provide a cross subgroup between ethnicity and economic disadvantageousness which was leveraged in this analysis. There may be additional benefits to adding in additional demographic fields. Perhaps the impact of the ethnic and economic subgroup is greater on of the genders? Or perhaps the impact is negated after a certain level of parental education? These important questions can only be accurately addressed with a completely individualized dataset. The privacy concerns and constraints by FERPA are not to be diminished, but perhaps this analysis could be done within the government confidentially and the summary of the findings published publicly.

Despite some of these limitations, the model was able to procure an R2 value of .88 meaning the collection of variables created was able to explain 88% of the variability in test scores. There are potentially better data points to stand in for neighborhood wealth and crime, but the variables created from disparate datasets were able to register as significant factors and contributed to the predictive capabilities of the model. The comparison between the initial model and final model illustrated the final model’s ability parse out some of the other factors such as neighborhood income that were hiding within the ethnic and economic subgroups of the initial model.

The R code used in the analysis has been uploaded to GitHub.com and can be found here: [https://github.com/mrcultrera/MIS500-1-Portfolio-Project](https://github.com/mrcultrera/MIS500-1-Portfolio-Project%20). The Tableau story and charts can be interacted with and found here: <https://public.tableau.com/views/PortfolioProjectEthnicityAdvantaged/Story1?:display_count=y&publish=yes&:origin=viz_share_link>

References

California Assessment of Student Performance and Progress. (2019). English language arts/literacy and mathematics [Data files]. Retrieved from <https://caaspp-elpac.cde.ca.gov/caaspp/ResearchFileList?ps=true&lstTestType=B&lstTestYear=2019&lstCounty=00&lstCntyNam=Select%20County...&lstFocus=btnApplySelections>

California Department of Public Health. (2020). Leading causes of death by zip code [Data file]. Retrieved from <https://data.ca.gov/dataset/leading-causes-of-death-by-zip-code>

California Department of Finance. (2020). ZIP code tabulations areas (ZCTA) in California: Table 1—total population, median age, sex, age, household and family size [Data file]. Retrieved from <http://www.dof.ca.gov/Reports/Demographic_Reports/Census_2010/>

Davies, T. (2016). *The book of R.* San Francisco, CA: No Starch Press, Inc.

Dietrich, D., Heller, B., & Yang, B. (2015*). Data science & big data analytics: Discovering, analyzing, visualizing and presenting data*. Indianapolis, Indiana: John Wiley & Sons.

Frost, J. (2020). Multicollinearity in regression analysis: problems, detection, and solutions. Retrieved from <https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/>

Garavaglia, S., Dun, A., & Hill, B.M. (1998). *A smart guide to dummy variables: four applications and a macro*.

Ghosh, B. (2017, September 29). Multicollinearity in R. Retrieved from <https://datascienceplus.com/multicollinearity-in-r/>

Shao, Q. (2016, September 24). Linear regression in R. Retrieved from <https://rpubs.com/AzureQ/lrinr>

State of California Franchise Tax Board. (2017). Personal income tax statistics by zip code [Data file]. Retrieved from <https://data.ca.gov/dataset/personal-income-tax-statistics-by-zip-code>