Maggie Kahn

Business Analytics

Professor Levkoff

November 17, 2020

Analytics Project 2

**Executive Summary**

For this project, I took the tidy data that I organized from the last project and expanded on it by creating 4 different linear regression models. The data that I used included the average sugar consumption in grams, GDP measured in billions of dollars, GDP per capita, and life expectancy in years over the span of 2006-2013. All of these variables were based upon a list of the same countries around the world. I had previously eliminated any countries that were missing data for the other variables so each would be the same. I took the data from my past project excel and transfer it to a new excel sheet. Within the workbook, I created multiple sheets to separate my data based on linear regression models that I would brainstorm later in the project. While I was brainstorming possible variables, I wanted to add a variable that represented the overall wealth of the country, so I chose to include poverty percentage as another variable in my project. In excel, I imported data for poverty level (%) across all of the countries I had previously accounted for. I then researched what the average poverty level for the world during the same time period (2006-2013). I decided to make this variable categorical so I could use it later on in my multiple linear regression. I created a new column in excel and used the IF statement to have all values above 12.7% register as “above the poverty line” and all values below the poverty line register as “below the poverty line”.

**Regression Models**

My next step was to brainstorm possible regression models that would produce helpful diagnostics. By just looking at my past graphs, I noticed that the United States had a very high sugar consumption and GDP. I also noticed that other countries who had a pretty large sugar consumption also had a high GDP. I hypothesized that countries that had a higher sugar consumption would also have higher GDP. So, I decided to make sugar consumption my main y-variable that would be tested with my other variables as well as be a part of the multiple linear regression. The first linear regression model was looking at sugar consumption for each country and its relation to life expectancy. I believe that countries with a higher sugar consumption would have a lower life expectancy. My second linear regression model was looking at sugar consumption in relation to GDP. My final linear regression model was looking at sugar consumption and its retaliation to GDP per capital. I hypothesized that people would have a shorter life expectancy than those countries with a smaller overall average sugar consumption. For my multiple linear regression, I wanted to examine sugar consumption in relation to GDP and the poverty percentage in each country. I hypothesized that countries that fell below the poverty level would show a smaller GDP as well as less sugar consumption.

**Diagnostic Results:**

**Regression #1:**

Equation:Y = -141.745 + 3.242x

My first regression was looking at sugar consumption (grams) in relation to life expectancy (years). I started with the training data partition. I previously predicted that there would be a negative correlation between sugar consumption (grams) and life. When I looked at the graph and the descriptive results, it showed a p-value extremely close one zero which indicates a significant relationship but unfortunately, it wasn’t in favor of my hypothesis. Instead of seeing a negative correlation, there was actually a positive correlation between life expectancy and sugar consumption. We can also look at the t-values for proof of significance. After looking at both the Intercept (beta 0) and the slope (beta 1), both are significant. The slope of the regression line for this graph is around 3.242, which signifies that each time there is a one unit increase in life expectancy, the sugar consumption variable is increased by 3.242 units. This proves that with an increase in sugar consumption, there was also an increase in life expectancy as well. For goodness of fit, the residual standard error is 28.87, which signifies each on average there is 28.87 worth of error on each data point. The r-squared value is 0.4908, which means about 49% of the information that I have collected can show why life expectancy is higher with higher sugar consumption. Overall, out of all the variables that I tested for, this is the relationship that performed the best. From looking at the histogram, I can see that the residuals are slightly skewed to the right but overall looks somewhat normal.

Chart, scatter chart

Description automatically generated

**Regression #2**

Equation: Y = 58.0551 + 0.9316x

The second regression I did included the relationship between sugar consumption and GDP (billions of dollars). After looking at the summary of the statistics, it looks completely different from the past graph. I originally hypothesized that higher sugar consumption contributed to a higher GDP. Immediately, I see that the intercept is 83.80, which shows that each time there is a unit increase in sugar consumption, there is also a unit increase in GDP of 83.80. When I look at the p-value, it shows statistical significance because both coefficients have a p-value below 0.05, so we can go ahead and reject the null hypothesis. Since the multiple r-squared is so small, I can tell that there are still a lot of variables that have to do with the GDP in each country. Looking at the residual standard error, I can see that there is a pretty large amount of error between each data point compared to the previous graph. When I looked at the histogram, I saw that this distribution is non-normal and extremely skewed to the left.

Chart, scatter chart

Description automatically generated

**Regression #3**

Equation: Y = 47.17 + 1.44x

This regression was quite interesting to look at. I wanted to see the relationship between sugar consumption and GDP per capita. The intercept for this graph was 47.17 units and the slope was 1.43 units. We say a larger difference with GDP and sugar consumption, but this is still quite a difference. After examining the p-value, you can tell that they are significant but they aren’t as small as some of the previous predictions. Similar to the previous regression, there is also quite a big residual standard error at 41.06. An multiple R-squared value of 0.04632 tells me that there are other variables besides sugar consumption that contribute to GDP per capita, which makes a lot of sense considering each country’s economic output. Looking at the histogram, the data looks somewhat normal but is also slightly skewed to the left.

Chart, scatter chart

Description automatically generated

**Regression #4**

Equation: Y = 89.65 + 0.7769x + -2.44

My multiple linear regression was looking at the linear regression relation between sugar consumption, poverty, and GDP per capita. The intercept for this graph was 33.60 units and the first slope for GDP per capita was 0.74 units. I used the poverty variable to judge the poverty percentage for each country compared to the sugar consumption. The slope for this variable was 45.21 units. Looking at the p-value for each variable, I noticed that the p-value for poverty percentage was statistically significant as well as sugar consumption, but the p-value for GDP per capita was not statistically significant. The residual standard error shows quite a bit of error with a value of 35.49 units. The multiple r-squared value was 29.6%, which means that my accounted for 29.6% of other variables that contributed to this relationship. The histogram shows the residuals have a higher frequency towards zero, which shows normality.

Chart, scatter chart

Description automatically generated

**Testing Data Partition**

My testing data showed very high results which is pretty unusual. The metric I used to partition my data was cubic because it showed the lowest output out of all the other methods. Sugar consumption and life expectancy performed the best out of both the training and testing data. I believe this was because sugar consumption and life expectancy showed the best p-value and a good multiple r-squared value.

**Conclusion**

To conclude, even though I didn’t initially believe that life expectancy and sugar consumption would have a correlation, it was interesting to see that it held the best results. I believe the cubic model was the best way to predict the variable of interest based off the analysis I did above.