Katrina Cokain

Project 3

For the multiple linear regression project, I chose to look at the different types of candy bars because I love all the different types. I wanted to see if certain aspects of a candy bar affected the number of calories. I decided to use a simple random sample by picking the aspects of a candy bar that are usually the looked at first by a person before they eat the candy bar. In this test, the dependent variables, or the variables that explain the number of calories in a candy bar, were sodium, sugar, protein, and fat calories. These variables were chosen because they contribute to the calories the most. The independent variable, or response variable that is affected because of the dependent variables, was the calories. For the data, I looked at the Data And Source Library, or DASL, website to get the data on candy bars and its ingredients. This source is a reliable source because it has a lot of different sets of collected data and has been used by many students for data collection. The data I used is shown in Figure 1.

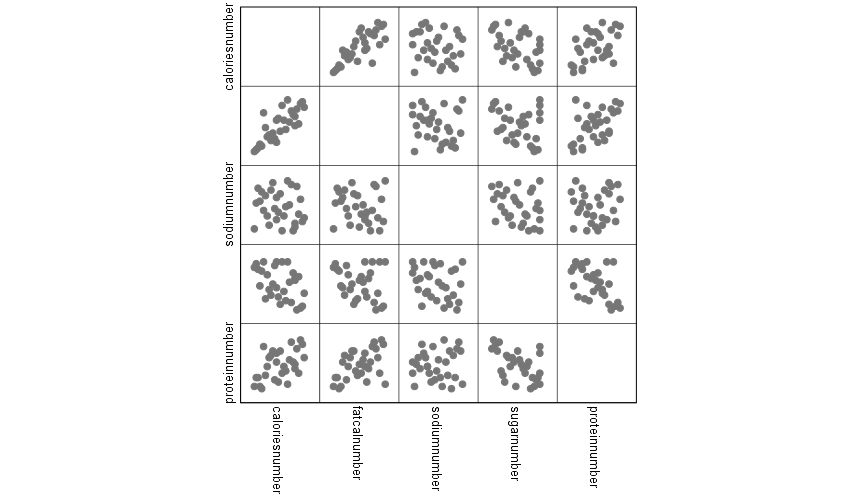
Figure 1





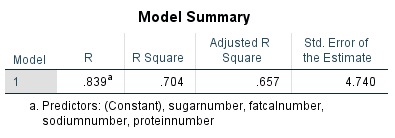
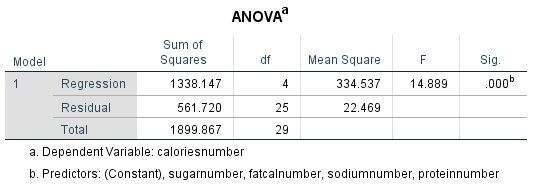
The hypotheses for this test are the null and alternative hypotheses. The null hypotheses is that all of the slopes are equal to zero. This means that the slopes are all equal together and if they all equal, then they are equal to zero. The alternative hypothesis is that at least one slope does not equal zero. This means that the slopes don’t equal each other, therefore, not equaling zero.

Before I could run a multiple linear regression test, I had to create scatterplots to see if the data shows there is a possible simple linear regression and if there is any positive linear relationships between the data. In Figure 2, the scatterplots show that it was not a simple linear regression. The scatterplots also show a possible linear relationship between calories and fat calories. I assumed that this simple random sample represented all candy bars.

Figure 2

For the test, I used a significance level of 0.01 because I thought that with the low sample size, the p-value would be lower than the significance level. After I determined by significance level, I continued with the test. I inputted my data into the computer program, SPSS, and performed the multiple linear regression. As shown in Figure 3, my r value was 0.839 and my r squared was 0.704. My F value and my p-value was 14.889 and 0.000, also shown in Figure 3.

Figure 3



To get my necessary data for my linear regression line, I needed to look at the coefficients of each variable. In Figure 4, the coefficients of each variable is shown in the unstandardized column. Each variable corresponds to a specific letter, x = sodium, z = sugar, w = protein, and t = fat calories. From the data, I created a linear regression equation of a line, which was y = -0.155x + (-0.289z) + 0.020w + 0.658t + 11.141. Before I determined if I had to reduce the variables to get a better linear regression line, I graphed the residuals of the data. Shown in Figure 5, the graphed residuals show a chaotic spread, both positive and negative with no real patterns. This means that it is possibly a good scatterplot.

Figure 4

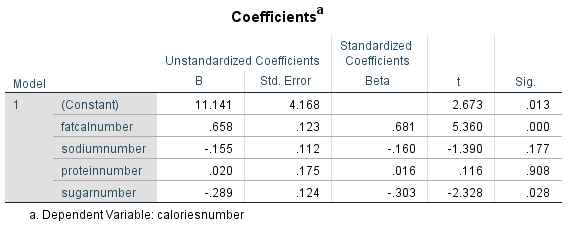
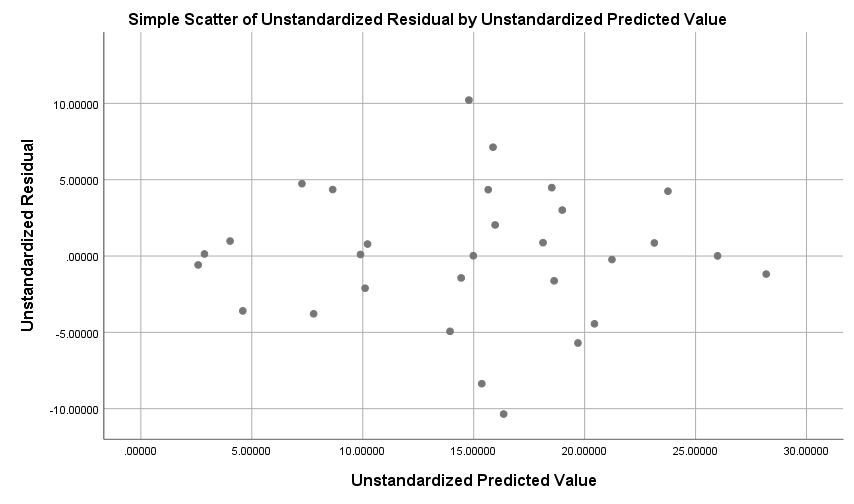


Figure 5



By looking at each variables’ significance value, I determined that I should remove protein from the test because its significance value was the largest, and I wanted to see if I could get the variables’ significance values below the significance level of 0.01. I ran the test again, this time without protein, and the r value, r squared value, and the p-value did not change, but the F value did change, shown in Figure 6. With the new test without protein, the unstandardized values and the significance values of each variable changed, which can be seen in Figure 7.

Figure 6

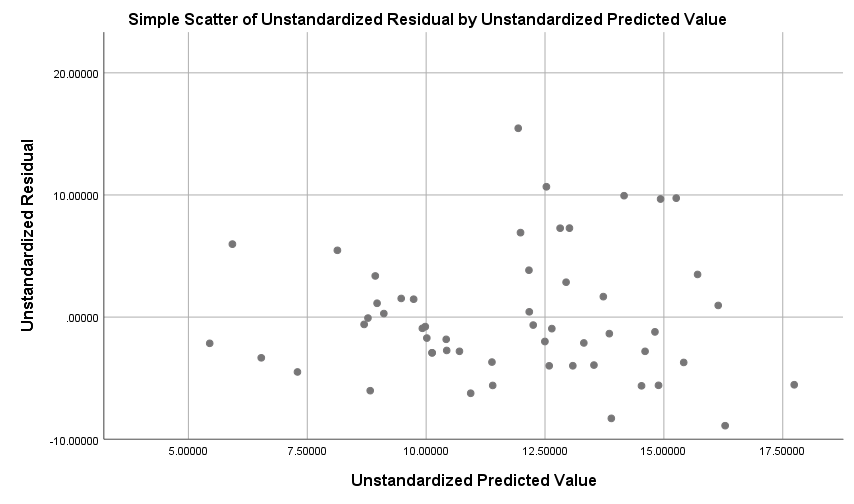
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ANOVAa** | | | | | | |
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 1337.842 | 3 | 445.947 | 20.630 | .000b |
| Residual | 562.024 | 26 | 21.616 |  |  |
| Total | 1899.867 | 29 |  |  |  |
| a. Dependent Variable: caloriesnumber | | | | | | |
| b. Predictors: (Constant), sugarnumber, fatcalnumber, sodiumnumber | | | | | | |

Figure 7

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Coefficientsa** | | | | | | |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
| B | Std. Error | Beta |
| 1 | (Constant) | 11.374 | 3.587 |  | 3.171 | .004 |
| fatcalnumber | .664 | .109 | .687 | 6.098 | .000 |
| sodiumnumber | -.155 | .109 | -.160 | -1.415 | .169 |
| sugarnumber | -.295 | .110 | -.310 | -2.673 | .013 |
| a. Dependent Variable: caloriesnumber | | | | | | |

The scatterplot of the residuals also changed, shown in Figure 8, when I removed protein from test. The residual scatterplot is still chaotic but there may be some outliers. Since the unstandardized values changed, the linear regression equation of a line changed, making it y = -0.155x + (-0.295z) + 0.664t + 11.374.

Figure 8



Again, I wanted to see if I could get the variables’ significance values lower than the significance level. The next largest significance value was sodium, so I decided to remove sodium from the test. Running the test again, the r value and the r squared value changed as shown in Figure 9. Also, the F value changed, in Figure 10, while the p-value stayed the same.

Figure 9

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Summaryb** | | | | |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | .825a | .681 | .658 | 4.735 |
| a. Predictors: (Constant), sugarnumber, fatcalnumber | | | | |
| b. Dependent Variable: caloriesnumber | | | | |

Figure 10

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ANOVAa** | | | | | | |
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 1294.569 | 2 | 647.284 | 28.873 | .000b |
| Residual | 605.298 | 27 | 22.418 |  |  |
| Total | 1899.867 | 29 |  |  |  |
| a. Dependent Variable: caloriesnumber | | | | | | |
| b. Predictors: (Constant), sugarnumber, fatcalnumber | | | | | | |

Again, the scatterplot of the residuals changed, shown in Figure 11, but is still chaotic with less data because I removed two variables. The unstandardized values changed, as shown in Figure 12, which changed the linear regression equation of a line to y = -0.247z + 0.696t + 8.005. This is the final linear function that I wanted to keep because even though the sugar’s significance value is higher than 0.01, the fat calories’ value and the constant’s value is lower than 0.01. This linear regression line is a good fit because more than half of the variables’ significance values are less that the significance level of 0.01. This proves that there is significant evidence that at least one coefficient is not zero.

Figure 11

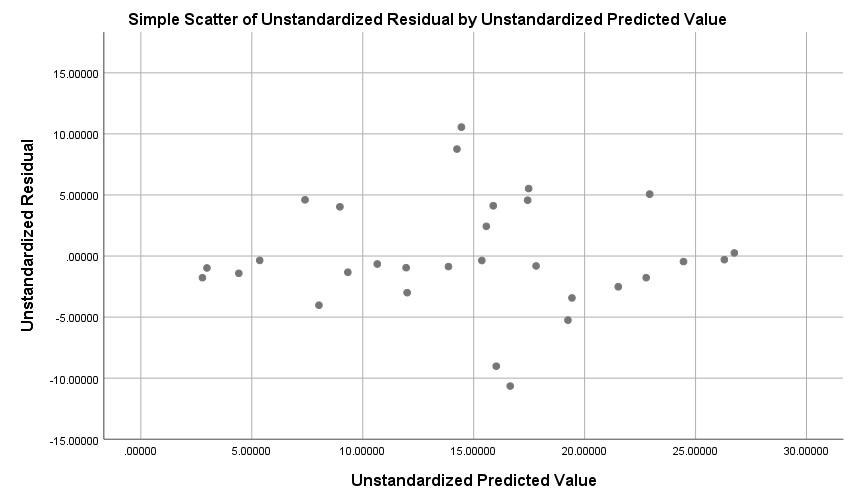


Figure 12

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Coefficientsa** | | | | | | |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
| B | Std. Error | Beta |
| 1 | (Constant) | 8.005 | 2.732 |  | 2.930 | .007 |
| fatcalnumber | .696 | .108 | .721 | 6.418 | .000 |
| sugarnumber | -.247 | .107 | -.259 | -2.308 | .029 |
| a. Dependent Variable: caloriesnumber | | | | | | |

In conclusion, the multiple linear regression test proved that there is significant evidence that at least one slope does not equal zero. It also proved that the amount of fat calories and sugars affect the number of calories in a candy bar.

Works Cited

“Nothing Found.” *DASL*, https://dasl.datadescription.com/datafiles/?\_sfm\_methods=Multiple Regression&\_sfm\_cases=4+59943.