Working Paper

What effects Weight the most: Using Interaction and Quadratic Regression to Find the Best Model

By William Chaudoin

For years I’ve had trouble with my Weight and my Waist size. I’ve underwent diets and strict gym regiments. I’ve accomplished many personal goals doing this, but it was always way more difficult than I desired it to be. Along the way I decided to keep track of my body weight and waist size, and what I consumed daily. That way I could try to track any patterns and then replicate them if they worked.

Meanwhile I found myself lacking confidence in my ability to code and perform statistical analysis. I wanted to refresh these skills and build an extensive portfolio using a variety of programs and languages. I decided to use random datasets and some I had a personal connection to. Believing that doing an analysis on a dataset that I was familiar with would help determine if I truly understood how to interpret and perform basic and advanced statistics or modeling.

I felt the best way to finish the creation of a portfolio was to use this tracked biomedical and dietary data that I routinely use in my study. To start with the raw data and go all the way through the process of creating multiple models, finding errors, seeking solutions, and just challenging myself to find if there were any answers.

This paper is not an academic paper seeking to be peer reviewed, it is solely evidence of my ability and skillset in this field but above all a personal challenge to myself. To make it more difficult for myself, I did not conduct this study with only one platform. This analysis was performed using Python within Jupyter Notebooks but also using R within the R Classic environment. The results for both platforms are featured throughout and in the appendix.

Starting this study, I did not know exactly what I would find nor was I entirely sure I was using the correct variables. However, that was by design. To truly understand a topic, one must be able to identify where they went wrong along the way. The end goal for this project is to go as far as the data will allow or that I am able.

I began this study by using linear regression models to find the best relationships between all variables with either my weight in pounds or my waist in inches. After doing this I was able to eliminate certain variables and narrow the focus to more significant relationships. Then I created several multivariate regression models to examine if multiple predictors worked better. This paper focuses on interaction and quadratic models with the goal of discovering if more complexity is an improvement. It utilizes what was learned from previous papers to construct new models.

**Data**

I collected 226 observations for 26 different variables. All quantitative variables include Weight in Pounds, Waist in Inches, Neck in Inches, Systolic Pressure in the Morning and at Night, Diastolic Pressure in the Morning and at Night, Body Temperature in the Morning and at Night, Pulse in the Morning and at Night, the amount of consumed Calories and Fat Calories per day, the amount of consumed grams of Fat, Sugar, Protein, and Fiber per day, and the total servings consumed per day.

Beyond these there are two qualitative variables measuring gym attendance and if cardio was performed, along with a Date variable. 4 remaining variables are Body Mass Index (BMI) calculated per day, a US Navy health metric called the Circumference Body Fat Index (CBF), the numbers of hours of Sleep per day, and the total intake in liters of Water per day.

I utilize all variables but Date in the study, thus using 25 of the 26 potential variables. They are all rounded to two decimal place whole numbers. Any missing values were replaced by the mean of that variable; Gym and Cardio are coded as dummy variables. For the multivariate models, I restricted the number of variables used to only Weight, Waist, Neck, Calories, Carbs, Fat, Protein, Fiber, and Sugar.

All values are observational data and not used for an experiment. A random sample from the overall sample was not taken, instead choosing to use all the data to train the models.

**Methodology**

9 ordinary least squares interaction and quadratic regression models were formed using the variables at my disposal with each model being built to predict Weight.

The purpose of these models was to find the most statistically significant relationship for Weight.

Several graphs were formed to look at the models with each variable combination. To judge the performance of each model, I tracked the Adjusted R-Squared, t-statistic, F-statistic, Pearson’s r value, Variance Inflation Factor, Durbin-Watson statistic, and Coefficient of Variation (COV).

Based on the results of the linear and multivariate regression models, I was able to eliminate all variables but Waist, Neck, Carbs, Protein, and Sugar.

**Models**

The equations for all models can be viewed below as Table 1 in the Appendix.

**Table 1: Models**

A screenshot of a computer

AI-generated content may be incorrect.

Model 1 is an interaction regression model that features Waist and Neck as the independent variables. Per Table 2 in the Appendix, it explains 79.4% of the variability in Weight, a very high amount. It is statistically significant with an F-Value of 290.80, far above the threshold of 2.6452. The model has a strong fit to the data as it has a low COV of 1.0578. The coefficients make reasonable sense although the intercept term has drastically ballooned into the thousands but considering zero is not within the range of the dependent variable it can be ignored. In comparison to the multivariate model using these terms, it is not a large improvement. Table 3 in the Appendix displays the model statistics for the multivariate models from the previous paper in this series. The Adjusted R-Squared for that model was only 78.3%; the inclusion of an interaction term increased that value by 1.1%. By including the interaction term, the statistical significance also decreased as the F-Value declined from 407.6637. Despite this, all coefficients are statistically significant based on a t-test, the values of which can be found in Tables 5 and 14 in the Appendix, so the interaction term does have some value in predicting Weight. There is no obvious lack of fit based on an analysis of the residual plot associated with this model, located in the Appendix in the Graphs 2 and 4 sections.

**Table 2: Model Statistics**

A screenshot of a table

AI-generated content may be incorrect.

Model 2 focuses on the relationship between Weight and the independent variables Carbs and Sugar. This was the best performing dietary model that utilized only 2 independent variables in the multivariate study. With the addition of the interaction term, it explains 10% of the variability in Weight. This is a decrease in the predictive power from the model lacking the interaction term, as it had an Adjusted R-Squared of 10.4%. The interaction model is statistically significant with an F-Statistic of 9.3690, above the needed 2.6452. This is also a decrease as the multivariate model had an F-Value of 14.0940. The multivariate model even has a lower COV as the interaction model received a COV of 2.2127 while the multivariate model received a 2.2080, suggesting greater variability in the calculation. As for the coefficients of the model, they largely do not make sense. The Sugar coefficient is negative, and the interaction term is roughly zero. Based on a t-test, the interaction term is not statistically significant; the values of which are in Tables 6 and 15. The Model 2 residual plot (located within Graphs 2 and 4 in the Appendix) also features several points beyond 2 standard deviations away from zero, suggesting a strong lack of fit with this model. Overall, an interaction model using Carbs and Sugar is not an improvement over the previous multivariate model in any way.

Model 3 also utilizes Carbs but drops Sugar in favor of using Protein as a predictor. This was a weaker model from the multivariate study that was statistically significant. Does the interaction term improve the predictive power? Based on Tables 2 and 3, the short answer to that is no. The Adjusted R-Squared for this model is 1.2%, lower than the previous model of 1.6%. It has an F-Value of 1.8730, meaning that it is not statistically significant. It has higher variability with a COV of 2.3195 than the multivariate model of 2.3144. Per the t-values in Tables 7 and 16, none of the coefficients are statistically significant. This is unsurprising as the interaction term is zero and both the Sugar and Protein coefficients are negative. The associated residual plot in Graphs 2 and 4 sections within the Appendix displays a strong lack of fit with several outliers.

Model 4 is the last interaction model. It explores whether the combination of Protein and Sugar is better than any interaction with Carbs. It is an improvement over the Carbs and Protein model but falls short of the Carbs and Sugar model. It also does not perform as well as the multivariate model using these same variables. It explains 4.5% of the variability in Weight, lower than the multivariate model of 4.7%. It has a statistically significant F-Value of 4.5220, which is also lower than the multivariate model of 6.5575 per Table 3. It also has a slightly higher COV of 2.2800, over the previous 2.2774. As for the associated coefficients, the interaction term like the previous 2 models is roughly zero. Both the Protein and Sugar terms are negative and much of the predictive power in this model is within the intercept term. As such I look at the t-values connected to the terms and find that none of them are statistically significant. The residual plot in Graphs 2 and 4 sections within the Appendix also displays a strong lack of fit with outliers. Overall, this model is not an improvement over the multivariate model in predicting Weight.

Model 5 is the first quadratic regression model of the bundle. Waist is the independent variable predicting Weight. Waist has consistently been a strong predictor as it was in the best performing linear regression and multivariate models without obvious issues. With a quadratic term included it explains almost as much of the variability as the interaction model featuring Waist and Neck. With 79.3% for the quadratic model compared to the previously mentioned 79.4% of the interaction model. It also has a higher F-Value with 431.90, meaning it not only is statistically significant but improves on the multivariate model of Waist and Neck. It also has a low COV of 1.0615, although this is slightly higher than the interaction model; there is some increased variation from including the quadratic term over the interaction term. The coefficients make logical sense outside of the inflated intercept term and all t-values are statistically significant. A scatterplot with a quadratic regression line was created for this model located in the Graphs 1 and 3 sections of the Appendix. It displays a strong relationship between the two variables with no obvious concerns. This is also true for the associated residual plot in the Graphs 2 and 4 sections.

Model 6 is another quadratic model using Neck to predict Weight. It is the third best model of all created in this bundle. It explains 30.5% of the variability in Weight, which is lower than the multivariate model with Waist and Neck, the Waist quadratic model, and the interaction model featuring Neck and Waist. Despite this it is still statistically significant with an F-Statistic of 50.2800. It has a COV that is still well below 10 but higher than other models with 1.9454, suggesting some increased variation and less fit. The coefficients make logical sense but there is an issue with the flipped sign for the quadratic term. Even more problematic, none of the coefficients are statistically significant. The associated scatterplot in the Graphs 1 and 3 sections of the Appendix also displays a lack of fit, with the line having no quadratic functionality. The residual plot in the Graphs 2 and 4 sections within the Appendix also displays a strong lack of fit with outliers. Despite being statistically significant, there are several issues with the ability to predict Weight with this model.

Model 7 uses Carbs in a quadratic model. It explains 1.8% of the variability in Weight, which is an improvement over the interaction model using Carbs and Protein but not even close to the ability of the Carbs and Sugar interaction model. It has an F-Value of 3.0050, which is lower than the threshold of 3.036 for statistical significance. It also has the second highest COV so far with 2.3124. The coefficients attached are roughly zero overall and none are statistically significant based on their t-values found in Tables 11 and 20. There is strong evidence for a lack of fit based on the associated scatter and residual plots located within the Graphs 1, 2, 3, and 4 sections of the Appendix.

Model 8 is a quadratic model using Sugar. It is an improvement on Carbs as it has an Adjusted R-Squared of 4.8%, ranking higher than all models using the dietary variables other than the interaction model between Carbs and Sugar. Unlike the Carbs quadratic model, the Sugar quadratic model is statistically significant. It has an F-Value of 6.7040, higher than the threshold of 3.036 for statistical significance. It also has the lowest COV value of any dietary model besides Model 2 with 2.2759. Despite these results, the coefficients attached are roughly zero overall. The quadratic term is not statistically significant based on the t-values found in Tables 12 and 21, but the term attached to Sugar solely is barely significant. There is strong evidence for a lack of fit based on the associated scatter and residual plots located within the Graphs 1, 2, 3, and 4 sections of the Appendix.

The final model of the bundle is Model 9 and it uses Protein in a quadratic regression. Unlike the Sugar quadratic regression, Protein is not statistically significant, nor does it explain a large value of the variability in Weight. It has an Adjusted R-Squared of 1.3% and an F-Value of 2.5220. It also received the second highest COV of all models with 2.3173. The coefficients attached are roughly zero overall and none are statistically significant based on their t-values found in Tables 13 and 22. There is strong evidence for a lack of fit based on the associated scatter and residual plots located within the Graphs 1, 2, 3, and 4 sections of the Appendix.

Beyond the test statistics and measurable values there are several issues with the coefficients of the models. Some models make sense as an equation while others have incorrect signs, or much of their predictive power is located within the intercept. The Waist and Neck models have the of their most predictive power based on their respective coefficients. Besides those, all models derive most of their power from the intercept and little value from the variable coefficients. The dietary models feature negative coefficients when predicting Weight. This makes no logical sense as it suggests eating more would make someone weigh less.

Due to the predictive issues surrounding these models, the concern of multicollinearity is present. Many of the T-statistics within these models for specific coefficients are not statistically significant despite the overall model being significant based on an F-Test. Between the dietary variables, there are significant high correlation parameters for each variable relationship which can be seen via the Correlation HeatMaps found in the Graphs 5 and 6 sections of the Appendix. Combining this with the negative coefficients there is strong evidence of multicollinearity.

I calculated Variance Inflation Factors (VIF) for each coefficient within each model which are in Table 4 of the Appendix. All models have high VIF values attached to the coefficients. It is safe to say there is a multicollinearity problem within these models, which is not surprising due to the nature of these models.

I also calculated Durbin-Watson statistics to determine if the residuals for each model were correlated. All statistics were less than 1, signifying they have a strong positive correlation. These values are also found in Table 4 below and in the Appendix.

**Table 4: Durbin-Watson and Variance Inflation Factor**

A screenshot of a table

AI-generated content may be incorrect.

**Conclusion**

Overall, Waist and Neck continue to be the best predictors of Weight. There are no interactions within the dietary variables that are significant. Sugar alone is a better predictor than either Carbs or Protein in a quadratic setting. Sugar works well with either Carbs or Protein in predicting Weight but no interaction is needed. In conclusion Waist is possibly the best predictor of all potential variables, regardless of model type.

Despite the multicollinearity issue, I feel something is wrong due to the negative coefficients associated with the dietary models and it is possible that I am using the wrong dependent variable. I feel a better dependent variable is the daily change in Weight. Daily consumption would have a larger role in the daily change rather than the long-term values. This is an idea I seek to explore in future studies.

Until then there are more models to explore and see if the relationship between my independent variable choices and the dependent variable of Weight improves or changes. Logarithmic, Cubic, and other types of models are future pathways of exploration. If those paths are fruitless then changing the dependent variable to daily change in Weight is a solution. There is also the idea of exploring Ridge Regression and Autoregressive models. In the end, I feel the best option will be to change the dependent variable but exploring all possibilities with the current setup interests me.

**Appendix: Tables**

**Table 1: All Interaction and Quadratic Models**

A screenshot of a computer

AI-generated content may be incorrect.

**Table 2: Interaction and Quadratic Model Test Statistics**

**A screenshot of a table

AI-generated content may be incorrect.**

**Table 3: Multivariate Model Statistics**

A table with numbers and lines

AI-generated content may be incorrect.

**Table 4: Durbin-Watson and Variance Inflation Factor**

**A screenshot of a table

AI-generated content may be incorrect.**

**Table 5: Model 1 (Python)**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Table 6: Model 2 (Python)**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Table 7: Model 3 (Python)**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Table 8: Model 4 (Python)**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Table 9: Model 5 (Python)**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Table 10: Model 6 (Python)**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Table 11: Model 7 (Python)**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Table 12: Model 8 (Python)**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Table 13: Model 9 (Python)**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Table 14: Model 1 (R)**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**Table 15: Model 2 (R)**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**Table 16: Model 3 (R)**

**A screenshot of a computer code

AI-generated content may be incorrect.**

**Table 17: Model 4 (R)**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**Table 18: Model 5 (R)**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**Table 19: Model 6 (R)**

**A screenshot of a computer code

AI-generated content may be incorrect.**

**Table 20: Model 7 (R)**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**Table 21: Model 8 (R)**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**Table 22: Model 9 (R)**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**Appendix: Graphs**

**Graphs 1: Scatterplots With Quadratic Regression Lines (Python)**

**Model 5**

**A graph with red dots and blue line

AI-generated content may be incorrect.**

**Model 6**

**A graph with red dots and a blue line

AI-generated content may be incorrect.**

**Model 7**

**A graph with red dots

AI-generated content may be incorrect.**

**Model 8**

**A graph with red dots

AI-generated content may be incorrect.**

**Model 9**

**A graph with red dots and blue line

AI-generated content may be incorrect.**

**Graphs 2: Residual Plots (Python)**

**A graph with blue dots

AI-generated content may be incorrect.**

**A diagram of a plot

AI-generated content may be incorrect.**

**A diagram of a plot

AI-generated content may be incorrect.**

**A diagram with blue dots

AI-generated content may be incorrect.**

**A graph of blue dots

AI-generated content may be incorrect.**

**A graph with blue dots and red lines

AI-generated content may be incorrect.**

**A diagram of a plot

AI-generated content may be incorrect.**

**A diagram of a plot

AI-generated content may be incorrect.**

**A diagram with blue dots

AI-generated content may be incorrect.**

**Graphs 3: Scatterplots With Quadratic Regression Lines (R)**

**A graph with a red line

AI-generated content may be incorrect.**

A graph with red line and numbers

AI-generated content may be incorrect.

A graph with a red line

AI-generated content may be incorrect.

A graph with a red line

AI-generated content may be incorrect.

A graph with a red line

AI-generated content may be incorrect.

**Graphs 4: Residual Plots (R)**

**A graph of orange and blue points

AI-generated content may be incorrect.**

**A graph of orange and blue points

AI-generated content may be incorrect.**

**A graph with orange dots

AI-generated content may be incorrect.**

**A graph with orange dots

AI-generated content may be incorrect.**

**A graph of orange and blue lines

AI-generated content may be incorrect.**

**A graph of a function

AI-generated content may be incorrect.**

**A graph with orange dots

AI-generated content may be incorrect.**

**A graph of orange and blue dots

AI-generated content may be incorrect.**

**A graph of orange and blue points

AI-generated content may be incorrect.**

**Graphs 5: Correlation Matrix Heatmap (Python)**

**A diagram of a food matrix

AI-generated content may be incorrect.**

**Graphs 6: Correlation Matrix Heatmap (R)**

**A diagram of a heatmap

AI-generated content may be incorrect.**