Working Paper

What effects Weight the most: Using Logarithmic and Cubic 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. I followed that by forming interaction and quadratic models with the goal of discovering if more complexity is an improvement. This paper utilizes what was learned from previous papers to construct various types of models, mainly a full second order model, a cubic model, and logarithmic 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.

This paper focuses only on the Weight, Waist, Neck, Carbs, Sugar, and Protein variables.

**Methodology**

7 ordinary least squares regression models were formed using the variables at my disposal with each model being built to predict Weight. One multivariate model using the best performing variables determined from past analysis. One polynomial regression model in the cubic style using the most consistent predictor from past analysis. A full second order model using the previous best two predictors. Finally, 4 logarithmic models using the 4 best predictors determined in the interaction and quadratic regression paper.

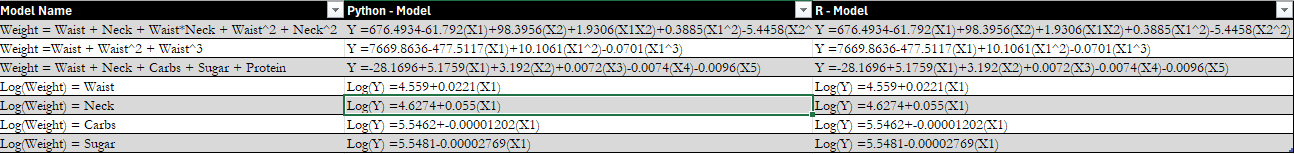
The purpose of these models is to find the most statistically significant relationship for Weight.

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, 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, multivariate, interaction, and quadratic regression models, I was able to eliminate all variables but Waist, Neck, Carbs, Protein, and Sugar as predictors.

**Models**

**Table 1: Models**



Model 1 is a full second order polynomial regression model using Waist and Neck to predict Weight. It includes Waist and Neck being used to predict Weight on their own as well as the respective quadratic terms for both and an interaction term. In the previous paper, these variables were the only ones that had a statistically significant model and a statistically significant interaction term. As such, I am curious to whether adding a quadratic term expands the accuracy. It explains 79.8% of the variability in Weight based on the Adjusted R-Squared term found in Table 2. This is a slight improvement over the interaction model, which explained 79.4% found in Table 5. It is statistically significant based on the F-Value of 179 and has a low COV of 1.048. The associated coefficients make logical sense despite the inflated values due to the nature of the model. However, the only statistically significant coefficient belongs to Waist. The interaction term moved from being significant to failing a t-test with the presence of the quadratic terms, which can be found in Tables 7 and 14. There is no evidence for a lack of fit based on the associated residual plots located in the Graph 5 and 16 sections of the Appendix, as less than 5% of the residuals are beyond 2 standard deviations.

I also performed a Nested F-Test to see if the simpler interaction model is superior to the full second order quadratic model. With an F-Statistic of 3.089 per Table 2, it narrowly surpasses the threshold of 3.037. Thus, the full order model passes the test and is superior to the interaction model.

**Table 2: Model Statistics**

**A screenshot of a computer

AI-generated content may be incorrect.**

Model 2 is a third order polynomial model using Waist as a predictor. It has 3 terms, where Waist is taken to the first, second, and third power. Waist was the only quadratic model in the previous paper which had a statistically significant quadratic term. As such, I wanted to explore whether including a cubic term would improve the predictive power. It does not, in fact it decreased the Adjusted R-Squared term from 79.3% in Table 5 to 79.2% in Table 2. It is still statistically significant based on the F-Value of 287.1. Despite explaining less of the variability in Weight, there is no giant increase in variation in the calculation as the model has a COV of 1.063; a slight increase over the quadratic model which had a COV of 1.062. The coefficients make logical sense, but none are statistically significant based on their corresponding t-statistics within Tables 8 and 15. There is no evidence for a lack of fit based on the associated residual plots located in the Graph 6 and 17 sections of the Appendix, as less than 5% of the residuals are beyond 2 standard deviations.

Model 3 is a standard multivariate model that includes the best predictor variables found during this study. Consistently, Waist and Neck were the best performing variables used in the model building process. Of the dietary variables, Carbs, Sugar, and Protein were the most statistically significant. I am curious if adding them all together would result in the best predictive model. That is not the case, but it did result in the best model featuring Carbs, Sugar, or Protein in predictive power. It has an Adjusted R-Squared of 78.9%, higher than any linear, multivariate, interaction, or polynomial model using these dietary variables per Tables 2, 3, 4, and 5. It is statistically significant per an F-Statistic of 169.5. It has one of the lowest COV values of any dietary model of 1.071, suggesting low variation. The good news ends there as the coefficients do not make logical sense as the terms associated with Sugar and Protein have flipped signs to negative. All terms are statistically significant besides Protein based on their t-values in Tables 9 and 16. There is no evidence for a lack of fit based on the associated residual plots located in the Graph 7 and 18 sections of the Appendix, as less than 5% of the residuals are beyond 2 standard deviations.

Model 4 is the first of the logarithmic models. 4 logarithmic models were formed to test the relationship between the best variables with a transformed dependent variable. This makes interpretation difficult as the R-Squared is not comparable to previous models due to this transformation. Nevertheless, Model 4 utilized Waist as the independent variable. It explains 77.6% of the variability in the log transformation of Weight per the R-Squared found in Table 2. It has a statistically significant value of 775.1 and a low COV of .198. The coefficients for this model are statistically significant per Tables 10 and 17 and represent that a 1-unit change in Waist results in a 2% change in Weight. There is no evidence for a lack of fit based on the associated scatter and residual plots located in the Graph 1, 8, 12, and 19 sections of the Appendix, as less than 5% of the residuals are beyond 2 standard deviations.

Model 5 is a logarithmic with Neck as the independent variable. It explains 31% of the variability in the log transformation of Weight. It is also statistically significant with an F-Value of 100.5. It also has a low COV of .347. The model coefficients are statistically significant per in Tables 11 and 18 and overall make logical sense as 5.5% change in Weight occurs with a 1-unit change in Neck size. There is strong evidence for a lack of fit based on the associated scatter and residual plots located in the Graph 2, 9, 13, and 20 sections of the Appendix, as more than 5% of the residuals are beyond 2 standard deviations.

Model 6 is another logarithmic model using Carbs to predict the log of Weight. Based on the associated R-Squared term, it explains 2% of the variability in the log transform of Weight. It is statistically significant based on the F-Value of 4.625 and has a low COV of .413. The model coefficients do not make logical sense as it is negative, meaning a 1-unit change in Carbs will decrease the log of Weight. The coefficient is also roughly zero. Despite this, the coefficient is statistically significant based on the t-value found in Table 12 and 19. There is strong evidence for a lack of fit based on the associated scatter and residual plots located in the Graph 3, 10, 14, and 21 sections of the Appendix, as more than 5% of the residuals are beyond 2 standard deviations.

Model 7 is the final logarithmic model, with Sugar as the independent variable. It is an improvement over the Carbs model by explaining 5.4% of the variability in the log transformation of Weight. It is statistically significant with an F-Value of 12.790 and has a low COV of .406. According to the t-values associated with the coefficients in Tables 13 and 20, they are statistically significant. There is strong evidence for a lack of fit based on the associated scatter and residual plots located in the Graph 4, 11, 15, and 22 sections of the Appendix, as more than 5% of the residuals are beyond 2 standard deviations.

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 Graphs 23 and 24 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 6 of the Appendix. All models but the logarithmic 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 their nature.

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 6 below and in the Appendix.

**Table 6: Durbin-Watson and Variance Inflation Factor**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Conclusion**

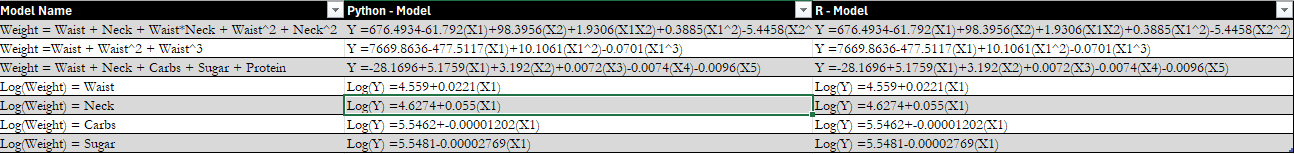
This study has been a time consuming and exhausting journey. I am not entirely sure what I have learned from these models. I have determined that Waist is probably the best overall predictor of Weight than any other variable. However, I feel the entire study is massively flawed. There are still methods to address this like creating log-log regression models or ridge regression models. Mainly I think the wrong dependent variable was chosen from the outset. Maybe not the entire study needs to be repeated but I should at least explore linear models in which my variables predict the daily change in Weight as the dependent variable.

With the daily change in Weight as the dependent variable, I feel that the issue surrounding the negative coefficients regarding the dietary variables will be addressed. Under the current models, they are trying to predict a value they have little influence over. In real life, the diet of a person day to day has more influence over the Weight of the next day rather than the Weight of the person next week. There is also the possibility of creating aggregate variables with the aggregate weekly consumption predicting the average weekly weight or something to that effect.

In conclusion, this study helped me practice statistical analysis and multiple coding languages. It forced me to ask questions and to find answers for a variety of issues. It allowed me to learn how to create models again and to practice the model building process. This will conclude this pathway of the study and force me to try a new method. In future installments I will use the daily change in Weight as the dependent variable, but I will also explore more advanced models. Attempting autoregressive models seems to be an interesting approach. I just mentioned that the daily consumption influences the daily Weight change, the same can be said for Weight as past Weight values should influence your future Weight. As such, using Weight to predict itself would be a good option. Overall, this journey has been rewarding and worth the investment.

**Appendix: Tables**

**Table 1: Logarithmic and Quadratic Models**

****

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

**A screenshot of a computer

AI-generated content may be incorrect.**

**Table 3: Linear Regression Models**

A screenshot of a table

AI-generated content may be incorrect.

**Table 4: Multivariate Model Statistics**

A table with numbers and lines

AI-generated content may be incorrect.

**Table 5: Interaction and Quadratic Models**

A screenshot of a computer

AI-generated content may be incorrect.

**Table 6: Durbin-Watson and Variance Inflation Factor**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Table 7: Model 1 (Python)**

A screenshot of a computer screen

AI-generated content may be incorrect.

**Table 8: Model 2 (Python)**

A screenshot of a computer

AI-generated content may be incorrect.

**Table 9: Model 3 (Python)**

A screenshot of a computer

AI-generated content may be incorrect.

**Table 10: Model 4 (Python)**

A screenshot of a computer

AI-generated content may be incorrect.

**Table 11: Model 5 (Python)**

A screenshot of a computer

AI-generated content may be incorrect.

**Table 12: Model 6 (Python)**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Table 13: Model 7 (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

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 code

AI-generated content may be incorrect.

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

A white and blue text with numbers and symbols

AI-generated content may be incorrect.

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

A screenshot of a computer code

AI-generated content may be incorrect.

**Appendix: Graphs**

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

**Graph 1: Model 4**

**A graph with blue dots and red line

AI-generated content may be incorrect.**

**Graph 2: Model 5**

**A graph with a red line

AI-generated content may be incorrect.**

**Graph 3: Model 6**

**A graph of a scatter plot

AI-generated content may be incorrect.**

**Graph 4: Model 7**

**A graph with blue dots and a red line

AI-generated content may be incorrect.**

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

**Graph 5: Model 1 Residual Plot**

**A graph with blue dots and a red line

AI-generated content may be incorrect.**

**Graph 6: Model 2 Residual Plot**

**A graph with blue dots

AI-generated content may be incorrect.**

**Graph 7: Model 3 Residual Plot**

**A diagram of a plot

AI-generated content may be incorrect.**

**Graph 8: Model 4 Residual Plot**

**A graph with blue dots and red line

AI-generated content may be incorrect.**

**Graph 9: Model 5 Residual Plot**

**A graph with blue dots and red lines

AI-generated content may be incorrect.**

**Graph 10: Model 6 Residual Plot**

**A diagram of a plot

AI-generated content may be incorrect.**

**Graph 11: Model 7 Residual Plot**

**A diagram of a plot

AI-generated content may be incorrect.**

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

**Graph 12: Model 4 Scatterplot**

**A graph with orange and blue lines

AI-generated content may be incorrect.**

**Graph 13: Model 5 Scatterplot**

**A graph with orange lines and numbers

AI-generated content may be incorrect.**

**Graph 14: Model 6 Scatterplot**

**A graph of a graph with numbers and a line

AI-generated content may be incorrect.**

**Graph 15: Model 7 Scatterplot**

**A graph with orange dots

AI-generated content may be incorrect.**

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

**Graph 16: Model 1 Residual Plot**

**A graph of orange and blue dots

AI-generated content may be incorrect.**

**Graph 17: Model 2 Residual Plot**

**A graph of orange and blue dots

AI-generated content may be incorrect.**

**Graph 18: Model 3 Residual Plot**

**A diagram of orange triangles

AI-generated content may be incorrect.**

**Graph 19: Model 4 Residual Plot**

**A graph of orange and blue lines

AI-generated content may be incorrect.**

**Graph 20: Model 5 Residual Plot**

**A graph with orange lines and numbers

AI-generated content may be incorrect.**

**Graph 21: Model 6 Residual Plot**

**A graph with orange dots

AI-generated content may be incorrect.**

**Graph 22: Model 7 Residual Plot**

**A graph with orange dots

AI-generated content may be incorrect.**

**Graph 23: Correlation Matrix Heatmap (Python)**

**A screenshot of a graph

AI-generated content may be incorrect.**

**Graph 24: Correlation Matrix Heatmap (R)**

**A diagram of a heatmap

AI-generated content may be incorrect.**