Working Paper

What effects Weight the most: Using Multivariate 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. This paper focuses on multivariate models with the goal of finding if adding variables is an improvement or a simpler model is better. Another desire is to see if dietary information is a good predictor of weight. Unlike the previous paper I undertook, this does not focus on predicting Waist, it instead narrows the focus to only weight.

**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**

26 ordinary least squares multivariate 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. Regarding dietary predictor models, I undertook a regression method where I added variables to create the best model until I used all potential variables. Then I chose the best model between those that used 2 predictors and up to 6 predictors.

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

Based on the results of the linear regression models, I was able to eliminate all variables but Waist, Neck, Carbs, Protein, and Sugar. Despite this, I decided to include some eliminated variables in this paper to create more models around consumption.

**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.

The highest R-Squared is associated with model that features Waist and Neck as independent variables. It also has the highest Adjusted R-Squared. This isn’t surprising as Waist and Neck were the best performing variables in the linear study; other than BMI which used Weight in its calculation. This model is statistically significant and has a COV of 1.0860, which is well below 10.

The remaining models were all dietary predictor models totaling 25. Since these models all use multiple predictors, I will focus on the Adjusted R-Square statistic over the standard R-Squared statistic but I will cover both.

The best model of the dietary predictor models utilizes Carbs, Sugar, Protein, and Fat as predictors. Based on Adjusted R-Squared, 14% of the variability in Weight is explained by this model. The R-Squared is 15.5% but that’s due to including more variables. This model is statistically significant with an F-Value of 10.1550, exceeding the threshold of 2.4124. It has a COV of 2.1635.

**Table 2: Model Statistics**

A table with numbers and lines

AI-generated content may be incorrect.

The remaining models are a combination of two or more predictors. Each provides different levels of statistical significance and accuracy. Of the overall 26 models, only 19 are statistically significant. Beyond the previous 2 models, 17 models remain that are statistically significant.

Only 8 models are statistically significant that utilize 2 predictors. The best performing of these uses Carbs and Sugar as predictors. The associated Adjusted R-Squared explains 10.4% of the variability in Weight. It has an F-Statistic of 14.0940 which is above the needed 3.0363. It has an R-Squared of 11.2% and a COV value of 2.2080.

The next best includes Sugar and adds Calories instead of Carbs. This is odd as the model that used Calories as a single predictor in my previous paper was not a statistically significant model. For some reason the combination of Sugar and Calories is statistically significant with an F-Value of 8.6963, above the needed 3.0363. However, it does not explain much of the variability of Weight as it has an Adjusted R-Squared of 6.4% and an R-Squared of 7.2%. It also has a COV of 2.2570.

The third best 2 predictor model continues to include Sugar and adds Fiber. It is not largely different from the previous model with an Adjusted R-Squared that explains 5.8% of the variability in Weight and an R-Squared that explains 6.6%. It has an F-Statistic of 7.9259, exceeding the threshold.

Sugar and Fat were used as predictors in a model that explained 5.5% of the variability in Weight for an Adjusted R-Squared. Unsurprisingly it has higher R-Squared at 6.3%. It has an F-Value of 7.5135 and a COV of 2.2682.

A model using Protein and Sugar explained 4.7% and 5.6% of the variability based on the Adjusted R-Squared and R-Squared associated. With an F-Value of 6.5575 it is statistically significant and a COV of 2.2774.

Only 3 models with 2 predictors remain that are statistically significant. These models are the first that utilize Carbs and Fiber as predictors, a second that feature Protein and Fiber, and the final has Calories and Fat. Carbs and Fiber has the best Adjusted R-Squared of the bunch with 2.2% of the variability explained in Weight. Protein and Fiber is next with 1.9% of the variability, while Calories and Fat is last with 1.8%. The R-Squared for each did not improve significantly, with 3.1%, 2.8%, and 2.7% respectively. The F-Values associated with each are 3.5599, 3.1767, and 3.0841. Calories and Fat barely surpassed the 3.0363 threshold.

All models that contained 3 or more variables are statistically significant. There are 4 models that contain only 3 variables. Based on the results of the models that contained only 2 variables, I determined that Sugar and Carbs were the best performing variables. Thus, I wished to determine what is the next best performing variable and the associated 3 variable model. The answer to this question is Protein. The combination of Carbs, Sugar, and Protein explains 13.9% of the variability in Weight per the Adjusted R-Squared and 15% per the R-Squared. It has an F-Value of 13.0624 and a COV of 2.1653.

The remaining 3 variable models all use Carbs and Sugar. The first includes Calories, the second features Fat, and the last has Fiber. The Calories model has an Adjusted R-Squared of 13.2% and an R-Squared of 14.4%. It has an F-Statistic of 12.4384 and a COV of 2.1731. The Fat model has an Adjusted R-Squared of 12.8% and an R-Squared of 13.9%. An F-Value of 11.9740 and a COV of 2.1790. Lastly the Fiber model explains 10% of the variability per Adjusted R-Squared and 11.2% per R-Squared. An F-Value of 9.3596 and a COV of 2.2129.

I created 3 models that have 4 independent variables. These feature Carbs, Sugar, and Protein based on the results of the models that contain only 3 variables. The best of these was previously mentioned above. It features Carbs, Sugar, Protein, and Fat. The next best includes Calories instead of Fat. It has an Adjusted R-Squared of 13.9% and an R-Squared of 15.4%. An associated F-Value of 10.0574 and a COV of 2.1651. The last has Fiber and explains 13.6% of the variability per the Adjusted R-Squared and 15.1% per the R-Squared. The F-Statistic is 9.8182 and the COV is 2.1691.

I only formed 3 more models, 2 of which had 5 independent variables and 1 included all 6 variables. The first I’ll discuss includes Carbs, Sugar, Protein, Fat, and Calories. It has an Adjusted R-Squared of 13.6% and an R-Squared of 15.5%, with an F-Value of 8.0976 and a COV of 2.1682. The second featured Carbs, Sugar, Protein, Fat, and Fiber. It has the same Adjusted R-Squared as the previous but an R-Squared of 15.1%. An F-Value of 8.0893 and a COV of 2.1684.

The final model includes all 6 variables of Carbs, Sugars, Protein, Fat, Calories, and Fiber. The associated Adjusted R-Squared of 13.2% and an R-Squared of 15.5%. An F-Value of 6.7176 and a COV of 2.1731.

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. Waist and Neck have the most predictive power based on coefficients. Besides that, 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. Combining this with the negative coefficients there is strong evidence of multicollinearity.

I calculated Variance Inflation Factors (VIF) for each coefficient within each model. Many of the models have high VIF values attached to the coefficients, and increases as more variables are added. It is safe to say there is a multicollinearity problem within these models.

I also calculated a Durbin-Watson statistic to determine if the residuals for each model were correlated. All statistics were less than 1, signifying they have a strong positive correlation.

**Table 3: Durbin-Watson and Variance Inflation Factor**

A screenshot of a computer

AI-generated content may be incorrect.

**Conclusion**

Overall, Waist and Neck continue to be the best predictors of Weight while Carbs, Sugar, and Protein is the most statistically significant model and the combination of Carbs, Sugar, Protein, and Fat explain the most variability in Weight of the dietary models.

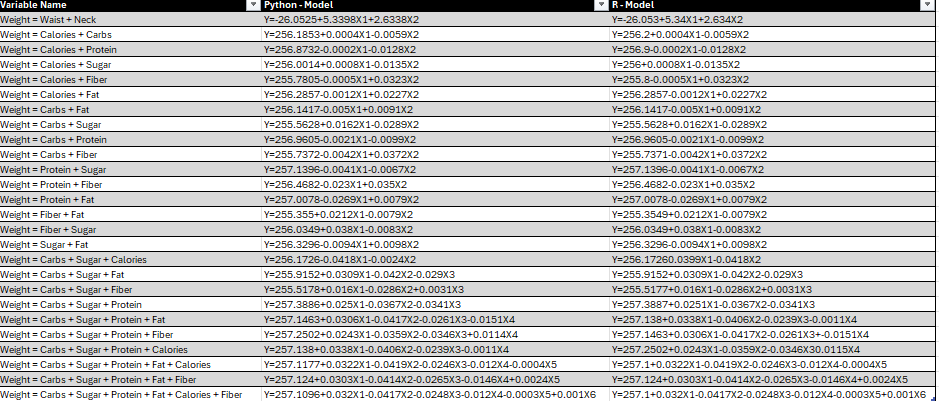
After analyzing the results of all these models, I have come to the stark conclusion that these variables do not really predict Weight. Adjusted R-Squared and R-Squared for most models is too low. Barely able to predict variability within double digits in half of the models. 19 of all models are statistically significant, leaving 7 models irrelevant. The COV for all models is well below the desired threshold of 10.

With the negative coefficients I feel something is wrong and I’m considering that maybe I’m using the wrong dependent variable. The multicollinearity problem does not make matters better, but I feel a better dependent variable is the 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. Quadratic, Interaction, 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 the next solution.

**Appendix: Tables**

**Table 1: All Models**

****

**Table 2: Model Test Statistics**

**A table with numbers and lines

AI-generated content may be incorrect.**

**Table 3: Durbin-Watson and Variance Inflation Factor**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Appendix: Graphs**

**Graph 1: 3D Scatterplots (Python)**

**A graph of a scatter plot

AI-generated content may be incorrect.**

**A graph with blue dots

AI-generated content may be incorrect.**

**A graph with blue dots

AI-generated content may be incorrect.**

**A graph with blue dots

AI-generated content may be incorrect.**

**A graph with blue dots

AI-generated content may be incorrect.**

**A graph with blue dots

AI-generated content may be incorrect.**

**A graph with blue dots

AI-generated content may be incorrect.**

**A graph with blue dots

AI-generated content may be incorrect.**

**A graph with blue dots

AI-generated content may be incorrect.**

**A graph with blue dots

AI-generated content may be incorrect.**

**A graph with blue dots

AI-generated content may be incorrect.**

**A graph with blue dots

AI-generated content may be incorrect.**

**A graph with blue dots

AI-generated content may be incorrect.**

**A graph with blue dots

AI-generated content may be incorrect.**

**A graph with blue dots

AI-generated content may be incorrect.**

**A graph with blue dots

AI-generated content may be incorrect.**

**Graph 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 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 of a plot

AI-generated content may be incorrect.**

**A diagram with blue dots

AI-generated content may be incorrect.**

**A diagram with blue dots

AI-generated content may be incorrect.**

**A diagram of a plot

AI-generated content may be incorrect.**

**A diagram with blue dots and red line

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 diagram of a plot

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 diagram with blue dots

AI-generated content may be incorrect.**

**A diagram 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 of a plot

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 diagram of a plot

AI-generated content may be incorrect.**

**Graph 3: 3D Scatterplots (R)**

**A graph of a diagram

AI-generated content may be incorrect.**

**A graph of calories and calories

AI-generated content may be incorrect.**

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

AI-generated content may be incorrect.**

**A graph of a graph with purple dots

AI-generated content may be incorrect.**

**A graph of a box with purple dots

AI-generated content may be incorrect.**

**A graph of a diagram

AI-generated content may be incorrect.**

**A graph of a graph with purple dots

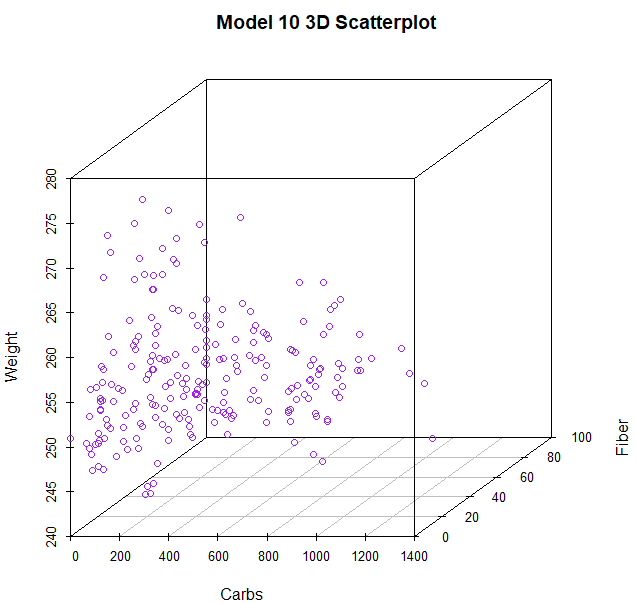
AI-generated content may be incorrect.**

**A graph of a box with purple dots

AI-generated content may be incorrect.**

**A graph of a diagram

AI-generated content may be incorrect.**

****

**A graph of a graph with purple dots

AI-generated content may be incorrect.**

**A graph of a model

AI-generated content may be incorrect.**

**A graph of a diagram

AI-generated content may be incorrect.**

**A graph of a graph with purple dots

AI-generated content may be incorrect.**

**A graph of a graph with purple dots

AI-generated content may be incorrect.**

**A graph of a model

AI-generated content may be incorrect.**

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

**A graph of a model

AI-generated content may be incorrect.**

**A graph with numbers and dots

AI-generated content may be incorrect.**

**A graph of a model

AI-generated content may be incorrect.**

**A graph of a model

AI-generated content may be incorrect.**

**A graph with numbers and dots

AI-generated content may be incorrect.**

**A graph with numbers and dots

AI-generated content may be incorrect.**

**A graph with numbers and dots

AI-generated content may be incorrect.**

**A graph with numbers and dots

AI-generated content may be incorrect.**

**A graph with numbers and dots

AI-generated content may be incorrect.**

**A graph with numbers and dots

AI-generated content may be incorrect.**

**A graph with numbers and dots

AI-generated content may be incorrect.**

**A graph with numbers and dots

AI-generated content may be incorrect.**

**A graph of a model

AI-generated content may be incorrect.**

**A graph with numbers and points

AI-generated content may be incorrect.**

**A graph with dots on a white background

AI-generated content may be incorrect.**

**A graph with dots on a white background

AI-generated content may be incorrect.**

**A graph with numbers and points

AI-generated content may be incorrect.**

**A graph with numbers and dots

AI-generated content may be incorrect.**

**A graph with numbers and dots

AI-generated content may be incorrect.**

**A graph of a model

AI-generated content may be incorrect.**

**A graph with numbers and dots

AI-generated content may be incorrect.**

**A graph with numbers and points

AI-generated content may be incorrect.**

**A graph with numbers and dots

AI-generated content may be incorrect.**

**A graph with numbers and dots

AI-generated content may be incorrect.**

**A graph with numbers and dots

AI-generated content may be incorrect.**

**A graph with numbers and dots

AI-generated content may be incorrect.**

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

**A close-up of a graph

AI-generated content may be incorrect.**

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

**A diagram of heatmap

AI-generated content may be incorrect.**