Predicting Calories
Abdullah Al Marzouq
Contact: almarzouk7@gmail.com

Introduction:

Calories are the singular most reliable metric of weight fluctuation when it comes to the human body. A calorie, in the case of this study, is the unit used to measure food energy, so by the same token, calories can be used to gauge a human's weight fluctuation. Since the 18th century, medical researchers throughout many institutions around the world have been researching the relationship between calories and human weight. The consensus, albeit disputed, is that one pound of fat is roughly 3,500 calories.

In the realm of food and nutrition, calories reign supreme. They're used as the header for all nutritional information in foods so, naturally, they play a huge role in the direction our weight fluctuates as humans. This is why a calorie deficit, eating less calories than what you burn, is a very popular method of weight loss. Although, it requires religious logging, that is, a person tracking their daily calorie consumption and expenditure on an hourly basis.

The US Food and Drug Administration (FDA) allows a margin of error of up to 20% and rarely audits products for calorie count. In addition to this, a study in the Journal of the American Medical Association finds that 19% of surveyed restaurants had deflated their calorie counts by a minimum of 100 calories. This poses an obstacle for a person committed to a caloric deficit because whether they're eating out or making a home cooked meal, their actual calorie consumption has a significant chance of being at least 20% higher than their calculated calorie consumption, no matter how scrutinous they are in logging.

On the other end of a caloric deficit, caloric expenditure (burning) is widely tracked using fitness watches. A series of independent studies set out to measure the accuracy of the top-selling watches in tracking caloric expenditure. The popular study subjects were the Apple Watch, WHOOP, and Fitbit. The results were a resounding error in each watch, with the lowest being the Apple Watch at 15% error and the highest being the Fitbit at 33%. It is noteworthy that other studies followed suit and found similar results.

I have personally been on a caloric deficit since eleven months ago, and as of today I have lost roughly 40lbs. However, I use a calorie tracking app and an Apple Watch to track my caloric deficit, and according to the data gathered from them, I should have lost 55lbs, a striking 33% error that inspired this project. I am fatigued from meticulously tracking my calories every hour of every day and would like to do it holistically without bearing the weight of errors. After all, weight loss has been present long before fitness watches and the FDA.

With all this in mind, I set out to answer two critical questions: **What boosts a person's calorie intake count the most?** and **What boosts a person's calorie burn the most?**

Data Summary:

To answer my two research questions, I found a dataset on Kaggle that fit my topic of my research the best. The dataset contains 54 variables and 20,000 observations, which is above the desirable amount needed to complete the research. The dataset, called Life Style Data, is a complete dataset of random self-reported daily lifestyle habits leaning towards a healthier lifestyle.

The response variable I'm looking for in my final model is a caloric deficit. If the response variable results in a negative number, that means the person has a caloric surplus and not a deficit, meaning they ate more than they burned. After reviewing the dataset in its entirety, I selected the following initial potential predictors based purely on intuition and eliminating redundant variables.

Variable	Type	Mean (SD)	Unit	Description
Protein	Continuous	95.3 (25.4)	grams	Daily protein consumption
Fat	Continuous	76.6 (20.7)	grams	Daily fat consumption
Carbohydrate	Continuous	250.8 (60.2)	grams	Daily carbohydrate consumption
Water	Continuous	2.4 (0.8)	liters	Avg daily water consumption
Sleep	Continuous	7.1 (1.2)	hours	Avg sleep per night
Exercise	Continuous	62.5 (35.7)	minutes	Daily exercise duration
Age	Continuous	34.6 (9.8)	years	Age of subject
BMI	Continuous	24.9 (3.5)	kg/m^2	Body mass index
Stress	Continuous (ordinal)	3.1 (1.0)	1-5 scale	Self-reported stress score
Workout	Categorical	Cardio 40%, Strength 35%, Yoga 15%, None 10%	-	Type of primary daily workout

Since intuition is not a reliable form of data screening and the dataset includes many caloric surpluses, I filtered the dataset to include only the observations with a caloric deficit and this resulted in **~2800 observations**, then found the best predictors for my final model using various screening methods explained in the next section:

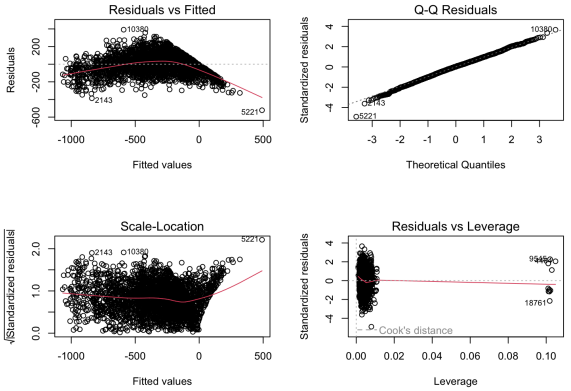
Variable	Type	Mean (SD)	Unit	Description
Session Duration	Continuous	1.6 (0.3)	hours	Daily session duration
Workout (4 Variables)	Categorical (Cardio dummy variable)	Cardio 40%, Strength 35%, Yoga 15%, None 10%	-	Type of primary daily workout
Weight	Continuous	59.1 (12.8)	kg	Subjects weight
Experience	Continuous (ordinal) (1=Beginner, 2=Intermediate, 3=Advanced)	2.4 (0.7)	1-3 scale	Fitness experience level
Workout Frequency	Continuous	3.9 (0.9)	days per week	Weekly workout frequency
Physical Exercise	Continuous	0.5 (1.0)	1-4 scale	Indicates the type or frequency of physical activity.
Protein	Continuous	99.5 (22.7)	grams	Daily protein consumption
Sugar	Continuous	24.1 (14.3)	grams	Daily sugar consumption
Age	Continuous	39.4 (11.7)	years	Age of subject

Methods and Results:

Three screening methods**, forward selection**, **backward elimination**, and **stepwise selection**, were used to identify the most relevant predictors of calorie deficit.

All three approaches showed clear agreement on a core group of predictors. *Session Duration, Workout Type, Weight, Experience Level, Workout Frequency, and Physical Exercise* were always retained, indicating they significantly contribute to the model. Forward and stepwise selection also included additional nutrition-related variables such as *Proteins, Sugar, and Age*, while backward elimination kept a slightly broader set because it started with all variables. All three screening methods produced similar AIC's, but Forward and Stepwise models achieved the lowest AIC values, suggesting the best balance of fit. Because the screening methods converged on similar predictors and the forward/stepwise models performed best by AIC, their shared variables, *Proteins, Sugar, and Age,* were selected for the final model.

In assessing the validity of the statistical procedures used and producing the final predictors, several core assumptions were evaluated. First, the assumptions underlying residual analysis were examined, including normality, linearity, and constant variance. **Normality** of residuals was evaluated using a Q-Q plot, which showed an approximately linear pattern with ever so slight deviations at the tails caused by a very small number of extreme observations. **Linearity** and **Constant Variance** were assessed through the residuals-versus-fitted plot, which revealed mild curvature and increasing residual spread at higher fitted values. These indicate some deviations from perfect linearity and constant variance; however, given the large sample size, these violations are not severe and do not compromise the model's usefulness. A leverage plot and **Cook's Distance** were used to assess influential points, and while several observations exhibited elevated influence, such as observations *5221, 9545, and 10380*, there was no evidence of heavy influence. **Studentized residuals** also identified four observations with absolute values above 3, marking them as outliers, but their influence was not large enough to warrant removal.



Second, the assumption of no **multicollinearity** was assessed using Variance Inflation Factors (VIF). This assumption requires predictors not to be excessively correlated, as this would inflate standard errors and destabilize coefficient estimates. Two predictors, *Proteins* and *Session Duration,* showed high multicollinearity, but were not manually removed because the stepwise screening methods had already filtered out redundant

variables. The remaining predictors each contributed significant explanatory power to the model. Additionally, robust regression confirmed that the final coefficients were stable, indicating that any remaining multicollinearity did not materially distort the results.

Back to influence, to further evaluate it on coefficients and predictions, **DFBETAS** and **DFFITS** were calculated. Several data points exceeded the cutoffs due to the large dataset size, which is expected when the model contains many predictors. These diagnostics confirmed that certain observations influenced specific coefficients, but no individual point heavily influenced the model as a whole. Thus, while the dataset contains some influential observations, they do not invalidate the fitted regression. To that end, I tested out robust regression for extra measure and to confirm the outliers are not influencing the final model.

To assess the influence of outliers on our final model, two **robust regression** models were fitted using Huber and Bisquare weighting and their coefficients were compared to the Ordinary Least Squares (OLS) method estimates. The table below summarizes the coefficients for all key predictors:

Variable	**OLS Coefficient**	**Huber Coefficient**	**Bisquare Coefficient**
Intercept (beta0)	719.213	728.279	728.621
Session Duration	-1143.434	-1150.451	-1151.167
Workout (HIIT)	-560.446	-557.557	-557.459
Workout (Strength)	-199.587	-198.139	-198.402
Workout (Yoga)	248.646	226.860	225.699
Weight	21.539	21.496	21.483
Experience Level	-124.703	-123.209	-123.456
Workout Frequency	44.858	45.121	45.423
Physical Exercise	12.469	12.230	11.994
Proteins	0.204	0.205	0.223
Sugars	-0.251	-0.241	-0.241
Age	-0.298	-0.322	-0.334

The coefficients are extremely similar across OLS, Huber, and Bisquare. For example, the effect of session duration on calorie balance is consistently large and negative, and the coefficients for Workout Type (HIIT and Strength), Weight, Experience Level, Workout Frequency, Physical exercise, Proteins, sugar, and Age all differ only in the second or third decimal place. The only visibly noticeable shift is for Yoga, where the estimated positive effect is slightly smaller under robust, but the direction and magnitude remain consistent across all three.

Overall, these results indicate that outliers and influential observations do not harm the relationships calculated by the OLS model. The robust coefficients closely follow the OLS estimates, therefore the OLS model is retained as our primary model, and the robust Huber and Bisquare fits were a sanity check that confirms the stability and reliability of our final predictor set.

No assumption violation was severe enough to justify alternative analytical methods. Although multicollinearity was high, the rest of the methods effectively addressed it. Mild deviations from normality and constant variance are not problematic in a large sample. Likewise, although several outliers and influential cases were identified, robust regression addressed them and concluded that they represent legitimate observations, and removing them would risk introducing bias. Therefore, no transformations, robust regression techniques, or alternative modeling strategies were required, and the final model remains appropriate and well-supported by the diagnostic checks.

Conclusions:

This analysis examined which factors most strongly influence calorie intake and calorie burn, using a large dataset and multiple linear regression methods. The results showed that *session duration, workout type, body weight, experience level, workout frequency, and physical exercise habits* are the most consistent and meaningful predictors of calorie deficit. Longer workout sessions and high-intensity workout types (particularly HIIT and Strength training) were associated with significantly larger calorie burn. On the intake side, dietary contributors such as *protein intake and sugar consumption* showed smaller and more stable effects on calorie deficits, meaning they boost intake the most out of all predictors. Together, the findings suggest that exercise-related behaviors have a substantially stronger impact on total calorie burn than dietary variables have on calorie intake, highlighting the dominant role of workout intensity and duration in shaping caloric deficits.

Several limitations should be acknowledged. The dataset relied on self-reported and/or device-estimated values, which may introduce measurement error, especially in calorie intake and burn estimates. Despite using robust regression to mitigate the effects of influential observations, some residual nonlinearity and nonconstant variance remain. Future work could explore nonlinear models, incorporate interaction effects between workout and dietary variables, or use time-series tracking to capture within-person changes over time. Additionally, a more detailed nutritional breakdown or integration with continuous heart-rate and metabolic data could further improve the accuracy of calorie deficit predictions.