Application of Multiple and Logistic Regression Within the Fast Food Industry

1 Introduction

Fast food culture within America is quite a popular phenomenon. Various chains such as McDonald's, Taco Bell, and Wendy's profit exponential amounts of money per year. For example, according to Macrotrends, "McDonald's annual gross profit was $13.207B, a 4.98% increase from 2021" (*Macrotrends*). Are the food items within these chains, however, representative of what is deemed "healthy" by our governing researchers and nutritionists?

According to a health article discussing this matter, they state that, on average, men should consume around 2,500 calories per day whereas women should consume around 2,000 calories per day (*NHS*). With this in mind, a generalization of the average amount of calories men and women should eat in a day is 2,250 calories. This could then be simplistically generalized to say that on average, one should eat around 750 calories within one of their three main meals of the day (breakfast, lunch, and dinner).

Within this report statistical analysis of a data set titled, "Fastfood", will be conducted. This dataset comes from the source titled, *OpenIntro*. The dataset consists of 515 observations across 17 variables. There is a restaurant variable that consists of 8 fast food chains: Mcdonalds, Chick Fil-A, Sonic, Arbys, Burger King, Dairy Queen, Subway, and Taco Bell. This attribute is then followed by a "food item" variable that represents the title of a various food item within a company. Finally, there are 15 nutritional indicators. Some of which are calories, trans fat, sugar, protein, cholesterol, fiber, etc. The data consists of no missing values but is noted to have some class imbalance. For example, Taco Bell has 115 food items recorded, whereas Chick Fil-A only has 27.

Two primary research questions will attempt to be answered:

1) Can the nutritional indicator, calories, be explained by a representation of other nutritional indicators that are independent of calories?

2) What are the leading fast food companies that contribute to a model predicting whether a fast food item will be unhealthy or not?

While attempting this, various regression models will be assessed. All assumptions will be considered.

2 Methodology

All methodology will be computed using the programming language *R Studio* (RStudio):

Method 1: Multiple Linear Regression

In trying to answer the first research question, multiple linear regression will be utilized. The general linear equation for this model is as follows:

$$\hat{Y} = \beta_0 + \beta_1(X_1) + ... + \beta_n(X_n) \quad (1)$$

Multiple linear regression is appropriate when trying to predict a quantitative dependent variable. In this case, the quantitative variable is the number of calories within a fast food item. According to a scholarly statistical textbook, the four primary assumptions that must be checked are linearity, errors being independently and identically distributed, errors having a normal distribution (referred to as the normality assumption), and errors having a constant variance.

Method 2: Binary Logistic Regression

Binary logistic regression, simplistically, has to do with modeling a binary dependent variable based on quantitative and/or qualitative explanatory variables. The primary mathematical concept encompassed within logistic regression is the logit. The logit is the logarithm of an odds ratio. The logit of Y is predicted by X. An interpretation of this can be noted by stating, "The logit is the natural logarithm (ln) of odds of Y, and odds are ratios of probabilities (π) of Y happening to probabilities (1–π) of Y not happening" (Peng). The multiple logistic model can be written as follows:

$$logit(Y) = natural\log(odds) = \frac{\pi}{1-\pi} = \alpha + \beta_1 X_1 + ... + \beta_n X_n \quad (2)$$

In this case, an additional binary indicator variable was added within the fast food data set. This attribute was labeled, *Unhealthy_Level* if a fast food item had a calories amount above the threshold of 750 calories. This threshold was decided upon based on the facts mentioned within the introduction of this report. The assumptions that will be checked within this model align with the primary assumptions noted within scholarly statistical textbooks. These include observation independence, absence of multicollinearity, linearity of independent variables and log odds, and a large sample size (Schreiber-Gregory).

3 Results and Discussion

Method 1: Multiple Linear Regression

Initial models are constructed to get a baseline feel for how the model predicts the number of calories within a food item. If all nutritional indicators are within the model, there are serious multicollinearity issues. This can be verified by seeing many of the predictors variance inflation factors (VIF'S) are above the threshold of 5:

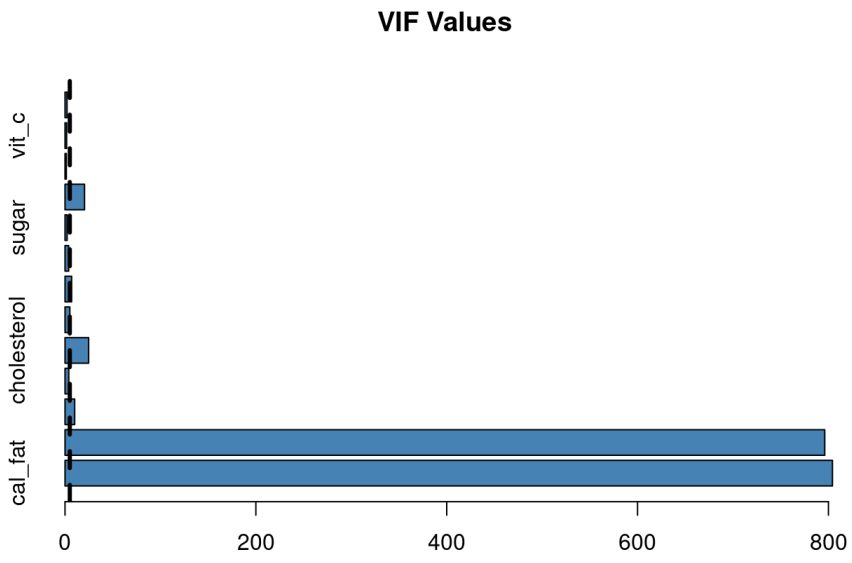

Figure 1. Full Model Variance Inflation Factors

Stepwise feature selection is conducted to reduce the number of nutritional predictors that are needed within the model for calories. It is found that an appropriate amount of predictors using this method is total fat, total carbs, cholesterol, saturated fat, trans fat, sodium, fiber, sugar, protein, vitamins A and C, and calcium. Although this is a subset of nutritional indicators, this did not fix VIF issues. Therefore, an adequate subset of predictors that was decided upon was calories regressed on total fat, total carbs, and cholesterol. The VIF'S for this model can be seen below and validate the multicollinearity assumption:

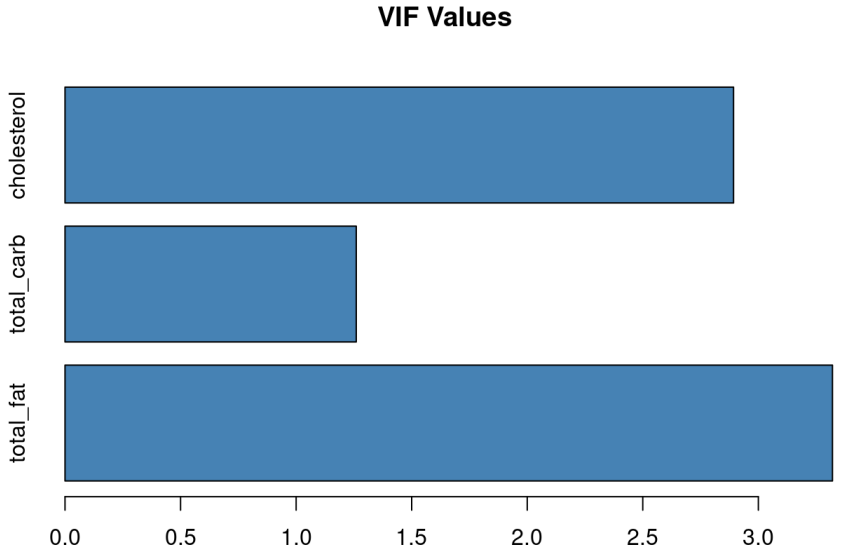

Figure 2. Reduced Model Variance Inflation Factors

This model has an adjusted R^2 value equal to 96.41%. That is, 96.41% of the variation in the number of calories within an item can be explained by the amount of total fat, carbohydrates, and cholesterol within the fast food item. Although this is a great percentage of variation being explained, assumptions must be assessed. It is found that almost every assumption is violated within the model and can be further validated within the plots below:

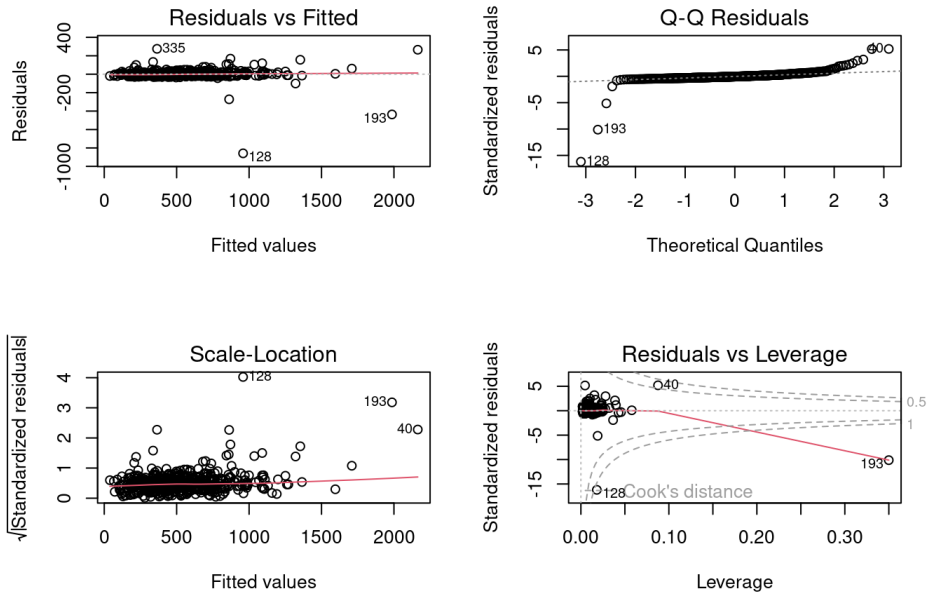

Figure 3. Multiple Linear Regression Assumption Plots

As seen, linearity is severely violated within the fitted vs. residual plot. The Q-Q plot allows one to see that normality of the standardized residuals is not appropriate/following the straight line. Independence of the errors can be conducted using the Durbin-Watson test. It is found that the p-value for this test is equal to .024. Therefore, there is evidence at the .2 alpha level that independence is not satisfied within this model. Constant variance can also be assessed within the residual vs. fitted or scale-point location plot which both show that constant variance is not satisfied.

Various transformations were conducted on the predictor and response variables. Such transformations can include taking the log of a variable, squaring the variable, taking the reciprocal, etc. This, however, did not improve assumption requirements. With this in mind, transitionary efforts were made to create a new response variable appropriate for logistic regression.

Method 2: Binary Logistic Regression

As mentioned, a new binary indicator variable was created to define which of the food items within the data set were deemed "unhealthy" with a calorie amount over 750. Logistic regression requires the splitting of the dataset into a testing and training set. The data

is partitioned and the restaurant variable within the data set is transformed into a factor attribute. A logistic model is created using the factored restaurant attribute. Classification is performed on the testing set and it is found that the accuracy of the model is 57.8%. The ROC Curve can be seen below. It is found that 65% of the area is under the curve:

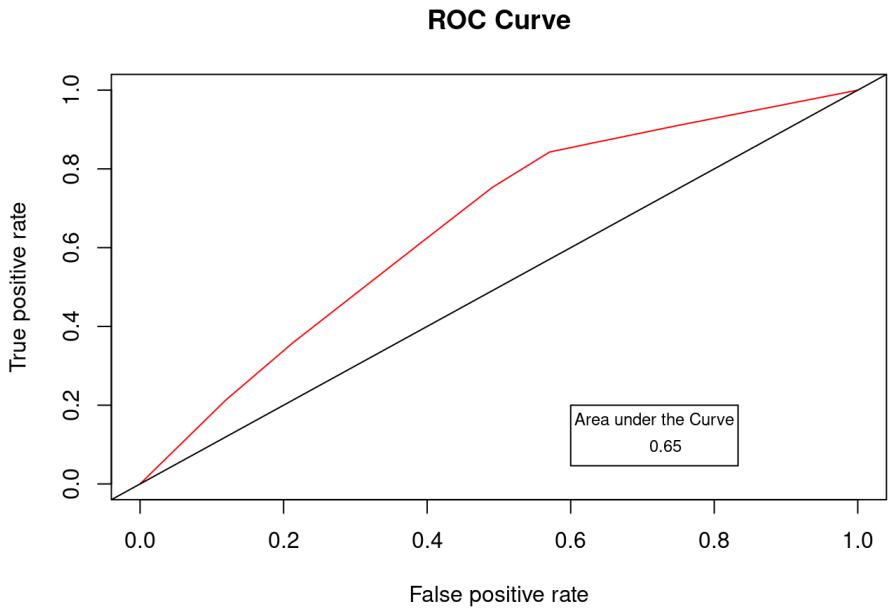

Figure 4. ROC Curve

To understand if there is a statistically significant difference between the unhealthiness level within various models, eight different logistic regression tests are conducted. One of the ANOVA (Analysis of Variance) tables is shown below:

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: unhealthylevel

Terms added sequentially (first to last)


                     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                                   514     474.13
as.factor(restaurant)  7   26.761       507     447.37 0.000368 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5. Coefficients Table for Reduced Model

As seen within figure 5, the p-value for the restaurant factor variable is $< .001$. Therefore, we have evidence at the .05 alpha level that there is a statistically significant difference between the restaurant that is used in predicting whether an item will be unhealthy or not.

Distribution in the amount of calories per restaurant can be seen below:

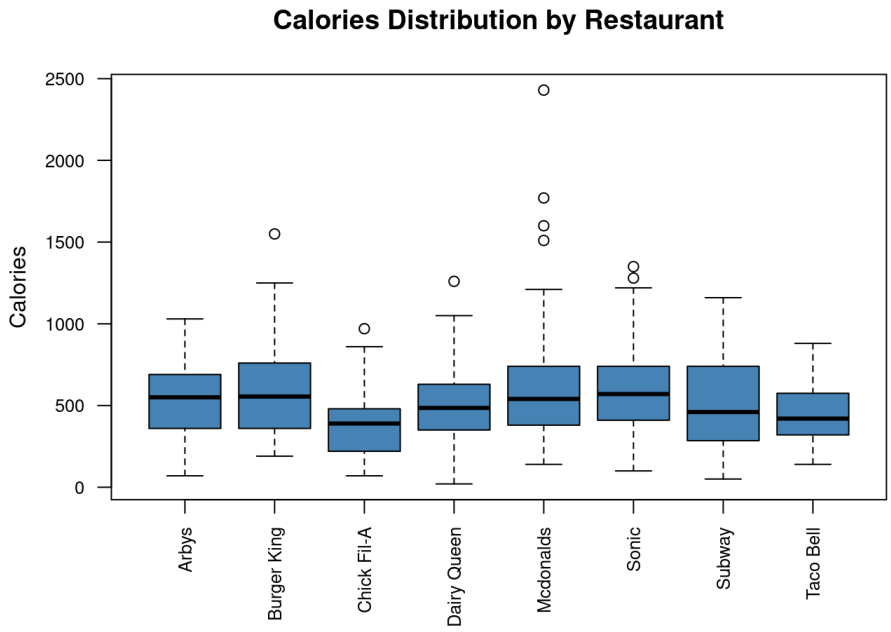

Figure 6. Distribution in Calories Per Restaurant

It is interesting to note that Mcdonald's and Subway have large inter quartile ranges. Chick Fil-A and Taco Bell seem to have the lowest median number of calories. Hypothesis' about these results pertaining to the logistic regression research question could include: Lower median calorie restaurants will have healthier outcomes within the model. Or, large IQR restaurants will have worse results within the model.

As mentioned, 8 different logistic regression models are conducted with the reference set to each subsequent restaurant. The odds ratios can be seen within the table below. The Y-axis points to the reference used and the X-axis points to the odds ratio that was computed:

	Mcdonalds	Taco Bell	Arby's	Chick Fil-A	Subway	Sonic	Burger King	DQ
Mcdonalds	1	.253** (-)	.265	.270	1.01	1.1	1.26	.796
Taco Bell	.395*** (+)	1	1.05** (+)	1.07	3.98	4.35*** (+)	4.98** (+)	3.14
Arby's	3.78	.953	1	1.02	3.79	4.14	4.75*** (+)	3
Chick Fil-A	3.69	.953	.98	1	3.72	4.06	4.66	2.941
Subway	.994	.251** (-)	.264	.269	1	1.09	1.253	.791
Sonic	.909	.230** (-)	.241	.246	.915	1	1.14	.724
Burger King	.793	.201*** (-)	.211*** (-)	.215*** (-)	.798	.872	1	.632
Dairy Queen	1.23	.318	.333	.3.4	1.26	1.38	1.58	1

Alpha Level:

*** : 0, ** : .001, * : .01.

+ : Column restaurant is unhealthier than reference

- : Column restaurant is healthier than reference

Figure 7. Odds Ratios Within Reference Models

Based on the results from figure 5, it is evident that the there is statistically significant differences in the prediction of the binary outcome unhealthy or healthy. Looking at Taco Bell as the reference, there is statistically significant evidence at the $< .0001$ alpha level that the odds ratio of an item being unhealthy at Taco bell was .395 times less probable than at Mcdonald's. In general, Taco Bell was statistically healthier than most other restaurants. Looking at Burger King, there tends to be healthier restaurants within this category. Insignificant restaurants that had similar outcomes within the models were Dairy Queen, Chick Fil-A, and Arby's (however it was significantly healthier than Burger King).

To finalize the validity of the various logistic regression models, assumptions were assessed. It is found that multicollinearity is not an issue due to the fact that all VIF's within each of the 8 reference models are under the threshold of 5.

It is found, however, that two of the primary assumptions within Logistic Regression are violated. For instance, independence is violated since the p-value within the Durbin-Watson test is equal to 0. That is, we have evidence at the .2 alpha level that independence is not satisfied within the mdoels. Goodness of fit within the models are also violated since the Hosmer-Lemeshow p-value is $< .0001$. That is, we have evidence at the .2 alpha level that the goodness of fit within the model is not satisfied.

This is why checking assumptions is important. Although the model classifies testing data moderately well, there are some discrepancies within the model pertaining to the overall fit and independence requirements. The pseudo R^2 within the model is also only equal to 5.6%.

4 Conclusion

Based on the findings within Multiple Linear Regression, it is interesting to note that total fat, carbohydrates, and cholesterol within a food item can explain the number of calories within an item quite well. It would be beneficial to test other nutritional indicator variables that are not as highly correlated with calories - or one another, to see if those results are maintained.

The findings within logistic regression were also noteworthy. It is interesting that there is a statistically significant difference of whether an item will be unhealthy or not based on the reference restaurant. In general, as stated, Taco Bell was the "healthiest" statistically significant restaurant. On average, restaurants that tended to have higher odds ratios of having an unhealthy food item were Burger King, Mcdonald's, and Subway (all at the .001 alpha level).

Although these results are interesting, statisticians and fast food companies should proceed with caution in making any food menu adjustments. More validation and testing needs to be done and various nutritional indicator variables that are not as highly correlated with calories or one another need to be assessed.

5 References

Fast Food Dataset., OpenIntro, www.openintro.org/data/index.php?data=fastfood.

Hadi, Ali S.. Regression Analysis by Example, John Wiley & Sons, Incorporated, 2012. ProQuest Ebook Central, https://ebookcentral-proquest-com.ezproxy.gvsu.edu/lib/gvsu/detail.action?docID=7103682.

McDonald's Gross Profit 2010-2023 — MCD, MacroTrends, 2023, www.macrotrends.net/stocks/charts/MCD/mcdonalds/gross-profit.

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/

Schreiber-Gregory, Deanna. Logistic and Linear Regression Assumptions: Violation Recognition and Control, 2018, www.lexjansen.com/wuss/2018/130_Final_Paper_PDF.pdf.

What should my daily intake of calories be?, NHS, 28 Sept. 2023