Stat108_FinalProject

Dylan Scoble and Aisha Lakshman

2/4/2022

Introduction

In the 2004 documentary Super Size Me, writer and director Morgan Spurlock took on a month-long challenge to only eat McDonalds food. Spurlock experienced a multitude of health issues, including weight gain, cholesterol spike, and negative impacts on his energy and mood, demonstrating the fast-food chains' instrumental role in America's obesity epidemic (Stossel 2006). Spurlock's film not only emphasized the consequences of caloric intake, but also brought light to the nutritional attributes of McDonalds menu items that caused adverse health effects. There are many factors that impact the quality and quantity of calories, such as levels of fat, protein, and carbohydrates, which is why many dieticians support the notion that "not all calories are created equal" (Tolar-Peterson, 2021). Spurlock's documentary and existing literature inspired an investigation of McDonalds menu items' caloric and nutritional records. Our research will address the following question: What nutritional attribute is most closely associated with calories for the McDonalds menu items? We will analyze a 2018 dataset from Kaggle titled "Nutritional Facts for McDonald's Menu" to answer our research question. Our chosen dataset provides nutritional information for all of McDonald's menu items, including calories, saturated fat, and cholesterol levels. Our research aims to guide the inspection of a nutritional label and to provide adequate information on the nutritional attributes that best estimates calories. Therefore, we will employ a modeling approach which estimates the closest association between calories and nutritional attributes. To see which nutritional attribute is the best estimator for calories, we will create a linear model of each nutritional attribute with calories as the response variable. A linear model for regression analysis is useful in answering our question because it will allow us to confidently determine what nutritional attributes hold the closest association to calories.

References

Tolar-Peterson, Terezie. 2021. "Not all calories are created equal - a dietician explains the different ways the kinds of foods you eat matter to your body". The Conversation. Retrieved Febuary 8th, 2022. https://theconversation.com/not-all-calories-are-equal-a-dietitian-explains-the-different-ways-the-kinds-of-foods-you-eat-matter-to-your-body-156900

Stossel, John. 2006. " 'Super Size Me' Carries Weight With Critics". ABC News. Retrieved Febuary 8th, 2022. https://docs.google.com/document/d/1XB-22QylvnbasBKe7n_DfkkENcZWRvIR5X8vZgOK6LY/edit#

Our Data

AISHA pls add the paragraph about how data was collected here

A sample of data can be viewed below.

Category	Item	Serving.Size	Calories	Calories.from.Fat	Total.Fat	Total.Fat....Daily.Value.	Saturated.Fat	Saturated.Fat....Daily.Value.	Trans.Fat	Cholesterol	Cholesterol....Daily.Value.	Sodium	Sodium....Daily.Value.	Carbohydrates	Carbohydrates....Daily.Value.	Dietary.Fiber	Dietary.Fiber....Daily.Value.	Sugars	Protein	Vitamin.A....Daily.Value.	Vitamin.C....Daily.Value.	Calcium....Daily.Value.	Iron....Daily.Value.
Breakfast	Egg McMuffin	4.8 oz (136 g)	300	120	13	20	5	25	0	260	87	750	31	31	10	4	17	3	17	10	0	25	15
Breakfast	Egg White Delight	4.8 oz (135 g)	250	70	8	12	3	15	0	25	8	770	32	30	10	4	17	3	18	6	0	25	8
Breakfast	Sausage McMuffin	3.9 oz (111 g)	370	200	23	35	8	42	0	45	15	780	33	29	10	4	17	2	14	8	0	25	10
Breakfast	Sausage McMuffin with Egg	5.7 oz (161 g)	450	250	28	43	10	52	0	285	95	860	36	30	10	4	17	2	21	15	0	30	15
Breakfast	Sausage McMuffin with Egg Whites	5.7 oz (161 g)	400	210	23	35	8	42	0	50	16	880	37	30	10	4	17	2	21	6	0	25	10
Breakfast	Steak & Egg McMuffin	6.5 oz (185 g)	430	210	23	36	9	46	1	300	100	960	40	31	10	4	18	3	26	15	2	30	20

Exploratory Data Analysis

We will also be removing all predictors that have "as % of Daily Value" attached at the end, since our purpose is not focused the daily values of the nutrients. These predictors add no value to our dataset or models.

The first thing we are doing is filtering out Total Fat (% Daily Value), Saturated Fat (% Daily Value), Cholesterol (% Daily Value), Sodium (% Daily Value), Carbohydrates (% Daily Value), Dietary Fiber (% Daily Value) from our nutritional attributes. These attributes don't aid in answering our research question, so we are taking these predictor variables out of consideration. However, we will keep in Vitamin A (% Daily Value), Vitamin C (% Daily Value), Calcium (% Daily Value), and Iron (% Daily Value) since our dataset records these attributes only in terms of daily value percentage.

In order to accurately compare the estimators, we want to standardize them first. To do this, we make sure every estimator is centered around a mean of zero with a variance of one. We create a new variable for each existing variable; this new variable contains the calculation of the formula $ new = (existing - mean(existing) / std_dev(existing)$

The dataset we are using contains a "Category" feature, which displays the type of food that McDonald's classifies the menu item as. We use this feature to divide our observations between food items (Breakfast, Beef & Pork, Chicken & Fish, Salads, Snacks & Sides, Desserts) and drink items (Coffee and Tea, Smoothies and Shakes, Beverages). The purpose of this is because we hypothesize that foods and beverages will have different best estimators. Most beverages do not have cholesterol or trans fats, so we think most of their calories will be tied to sugars.

A sample of the final version of our datasets (food and beverage) can be viewed below.

Food Data:

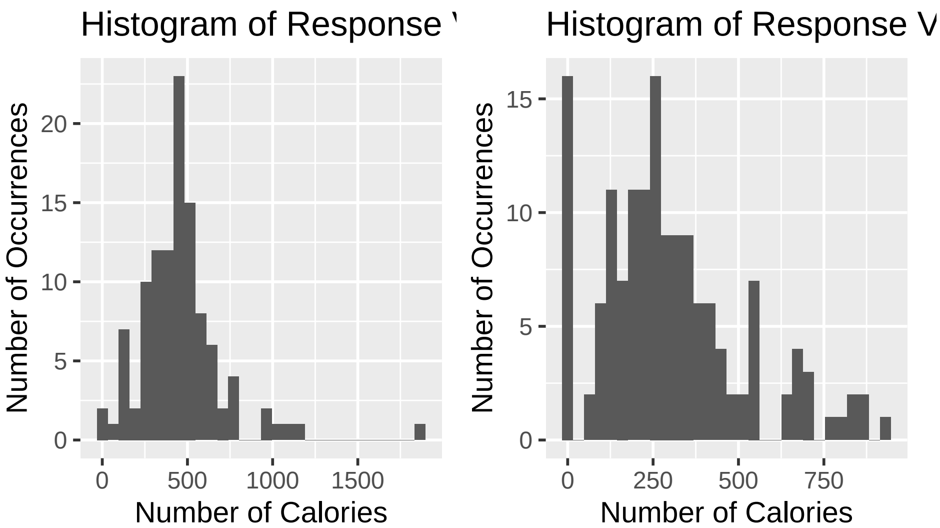
Category	Item	Serving.Size	Calories	Calories.from.Fat	Total.Fat	Saturated.Fat	Trans.Fat	Cholesterol	Sodium	Carbohydrates	Dietary.Fiber	Sugars	Protein	Vitamin.A....Daily.Value.	Vitamin.C....Daily.Value.	Calcium....Daily.Value.	Iron....Daily.Value.	Total.Fat.cent	Saturated.Fat.cent	Trans.Fat.cent	Cholesterol.cent	Sodium.cent	Carbohydrates.cent	Dietary.Fiber.cent	Sugars.cent	Protein.cent	Vitamin.A....Daily.Value.cent	Vitamin.C....Daily.Value.cent	Calcium....Daily.Value.cent	Iron....Daily.Value.cent
Breakfast	Egg McMuffin	4.8 oz (136 g)	300	120	13	5	0	260	750	31	4	3	17	10	0	25	15	-0.0820347	-0.1893492	-0.4750187	2.3497128	0.4406211	-0.5785792	1.511262	-0.9213133	0.3204526	-0.1406414	-0.3239491	0.2366001	0.8328746
Breakfast	Egg White Delight	4.8 oz (135 g)	250	70	8	3	0	25	770	30	4	3	18	6	0	25	8	-0.4339987	-0.5651567	-0.4750187	-0.3431026	0.4752816	-0.6139746	1.511262	-0.9213133	0.4079712	-0.3048021	-0.3239491	0.2366001	0.0304226
Breakfast	Sausage McMuffin	3.9 oz (111 g)	370	200	23	8	0	45	780	29	4	2	14	8	0	25	10	0.6218933	0.3743621	-0.4750187	-0.1139268	0.4926118	-0.6493701	1.511262	-0.9561810	0.0578969	-0.2227218	-0.3239491	0.2366001	0.2596946
Breakfast	Sausage McMuffin with Egg	5.7 oz (161 g)	450	250	28	10	0	285	860	30	4	2	21	15	0	30	15	0.9738573	0.7501696	-0.4750187	2.6361825	0.6312537	-0.6139746	1.511262	-0.9561810	0.6705269	0.0645593	-0.3239491	0.5303730	0.8328746
Breakfast	Sausage McMuffin with Egg Whites	5.7 oz (161 g)	400	210	23	8	0	50	880	30	4	2	21	6	0	25	10	0.6218933	0.3743621	-0.4750187	-0.0566329	0.6659142	-0.6139746	1.511262	-0.9561810	0.6705269	-0.3048021	-0.3239491	0.2366001	0.2596946
Breakfast	Steak & Egg McMuffin	6.5 oz (185 g)	430	210	23	9	1	300	960	31	4	3	26	15	2	30	20	0.6218933	0.5622659	1.8552618	2.8080644	0.8045560	-0.5785792	1.511262	-0.9213133	1.1081198	0.0645593	-0.2480350	0.5303730	1.4060546

Beverage Data:

Category	Item	Serving.Size	Calories	Calories.from.Fat	Total.Fat	Saturated.Fat	Trans.Fat	Cholesterol	Sodium	Carbohydrates	Dietary.Fiber	Sugars	Protein	Vitamin.A....Daily.Value.	Vitamin.C....Daily.Value.	Calcium....Daily.Value.	Iron....Daily.Value.	Total.Fat.cent	Saturated.Fat.cent	Trans.Fat.cent	Cholesterol.cent	Sodium.cent	Carbohydrates.cent	Dietary.Fiber.cent	Sugars.cent	Protein.cent	Vitamin.A....Daily.Value.cent	Vitamin.C....Daily.Value.cent	Calcium....Daily.Value.cent	Iron....Daily.Value.cent
Beverages	Coca-Cola Classic (Small)	16 fl oz cup	140	0	0	0	0	0	0	39	0	39	0	0	0	0	0	-0.9971411	-1.128868	-0.4750187	-0.6295723	-0.8591462	-0.2954157	-1.040219	0.3339258	-1.167363	-0.551043	-0.3239491	-1.232264	-0.8866654
Beverages	Coca-Cola Classic (Medium)	21 fl oz cup	200	0	0	0	0	0	5	55	0	55	0	0	0	0	0	-0.9971411	-1.128868	-0.4750187	-0.6295723	-0.8504811	0.2709112	-1.040219	0.8918098	-1.167363	-0.551043	-0.3239491	-1.232264	-0.8866654
Beverages	Coca-Cola Classic (Large)	30 fl oz cup	280	0	0	0	0	0	5	76	0	76	0	0	0	0	0	-0.9971411	-1.128868	-0.4750187	-0.6295723	-0.8504811	1.0142153	-1.040219	1.6240325	-1.167363	-0.551043	-0.3239491	-1.232264	-0.8866654
Beverages	Coca-Cola Classic (Child)	12 fl oz cup	100	0	0	0	0	0	0	28	0	28	0	0	0	0	0	-0.9971411	-1.128868	-0.4750187	-0.6295723	-0.8591462	-0.6847655	-1.040219	-0.0496195	-1.167363	-0.551043	-0.3239491	-1.232264	-0.8866654
Beverages	Diet Coke (Small)	16 fl oz cup	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	-0.9971411	-1.128868	-0.4750187	-0.6295723	-0.8418160	-1.6758377	-1.040219	-1.0259165	-1.167363	-0.551043	-0.3239491	-1.232264	-0.8866654
Beverages	Diet Coke (Medium)	21 fl oz cup	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	-0.9971411	-1.128868	-0.4750187	-0.6295723	-0.8244858	-1.6758377	-1.040219	-1.0259165	-1.167363	-0.551043	-0.3239491	-1.232264	-0.8866654

Response Variable

The next step is to create histograms for occurences of food items (food_data) and occurences of drink items (bev_data) against our response variable (calories). Based on the plots below, it seems that for both datasets, the Calories variable follows a normal distribution. ^ AISHA can u add a sentence why the normal distribution is a good thing



The summary statistics for calories (mean, median, standard deviation, IQR) for food items and drink items is also displayed below. Notice that for both datasets, the mean is larger than the median. This is likely attributed to the fact that there are many items that McDonald's advertises as "Zero Calories". The large group of zero calorie items acts as an outlier for our normal distribution and brings the median down without affecting the mean as much.

Food Data:

mean	median	std_dev	iqr
462.0909	445	249.3343	210

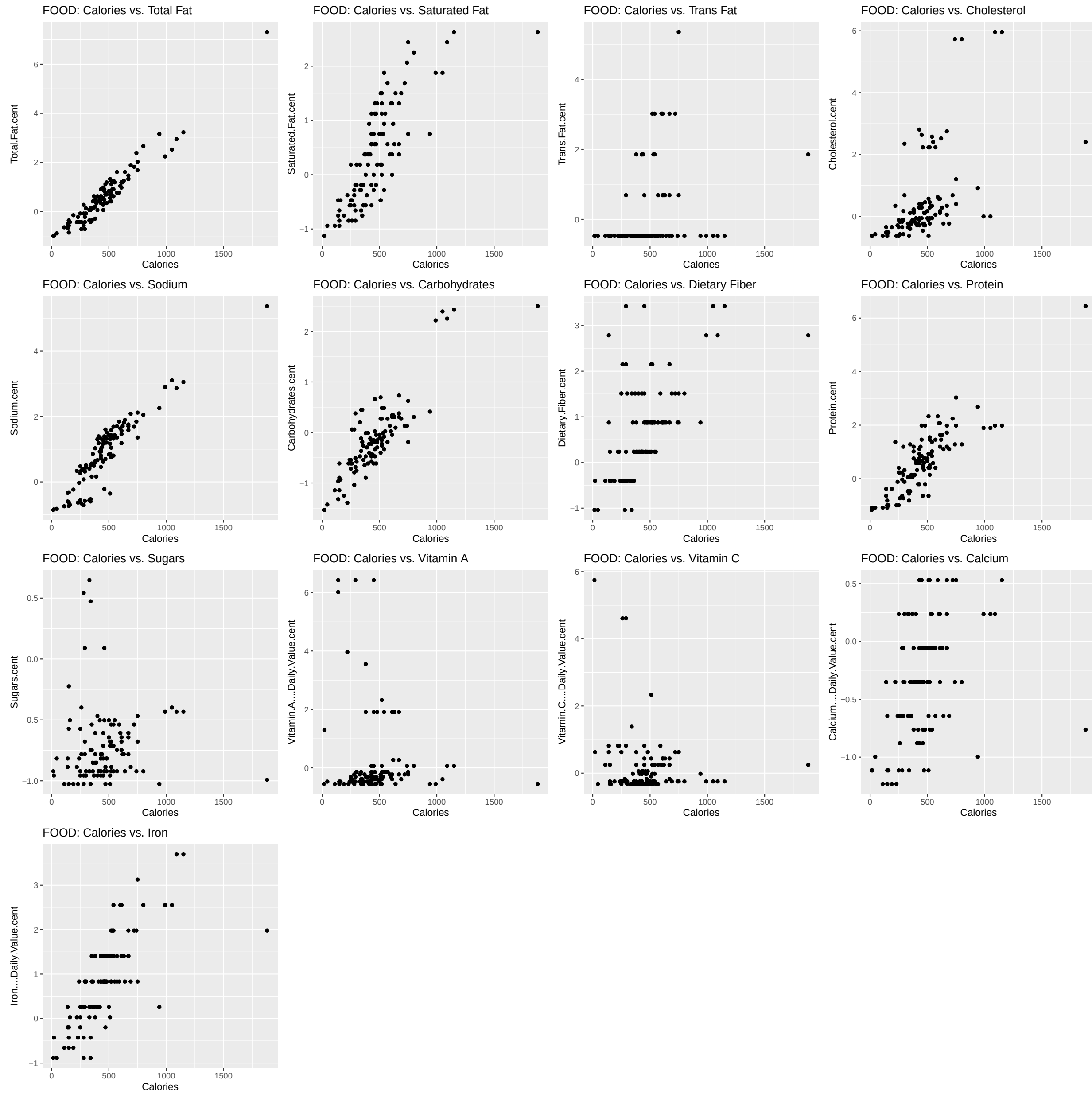
Beverage Data:

```
kable(t2, format = "markdown")
```

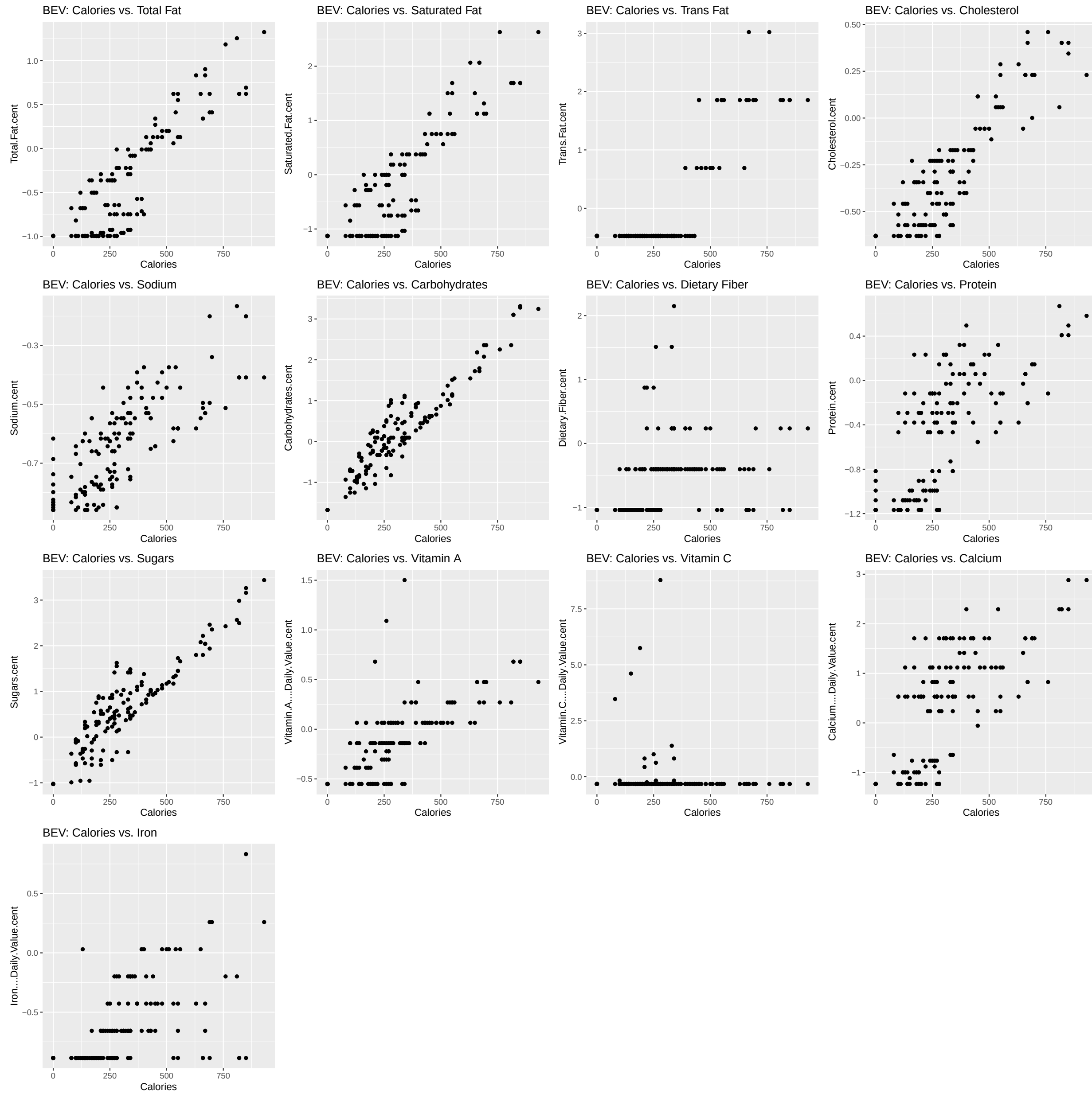
mean	median	std_dev	iqr
299.4667	270	208.8215	235

Estimator Variables

We also want to view a plot of each variable's relationship with our response. Below is every plot that describes the food dataset. ˆ AISHA can you add some here that talks about what the plots are showing and why we are showing them



Below is every plot that describes the beverage dataset. ˆ AISHA can you add some here that talks about what the plots are showing and why we are showing them



Model Selection with AIC

In order to conduct our experiment, we create a model that uses Calories as a response variable, and all centered variables as estimators.

The coefficients of our food model are displayed below.

term	estimate	std.error	statistic	p.value
(Intercept)	371.196	1.193	311.196	0.000
Total.Fat.cent	126.483	1.564	80.869	0.000
Saturated.Fat.cent	3.433	1.464	2.344	0.021
Trans.Fat.cent	0.563	0.656	0.858	0.393
Cholesterol.cent	-0.812	0.477	-1.703	0.092

term	estimate	std.error	statistic	p.value
Sodium.cent	-1.035	1.711	-0.605	0.547
Carbohydrates.cent	116.617	2.030	57.453	0.000
Dietary.Fiber.cent	-2.116	0.810	-2.614	0.010
Sugars.cent	-1.728	2.805	-0.616	0.539
Protein.cent	45.460	1.235	36.800	0.000
Vitamin.A....Daily.Value.cent	0.405	0.387	1.046	0.298
Vitamin.C....Daily.Value.cent	1.105	0.519	2.130	0.036
Calcium....Daily.Value.cent	-0.921	1.400	-0.657	0.512
Iron....Daily.Value.cent	-0.576	1.175	-0.490	0.625

In mathematical terms, we are creating an equation for a line where the x values are values of our estimators, and the y value is the response variable.

The equation for our food dataset can be read as the following:

$Calories = 371.196 + 126.483(Total.Fat) + 3.433(Saturated.Fat) + 0.563(Trans.Fat) - 0.812(Cholesterol) - 1.035(Sodium) + 116.617(Carbohydrates) - 2.116(Dietary.Fiber) - 1.728(Sugars) + 45.460(Protein) + 0.405(Vitamin.A....Daily.Value) + 1.105(Vitamin.C....Daily.Value) - 0.921(Calcium....Daily.Value) - 0.576(Iron....Daily.Value)$

The coeficients of our food model are displayed below.

term	estimate	std.error	statistic	p.value
(Intercept)	356.609	5.746	62.060	0.000
Total.Fat.cent	103.013	8.553	12.044	0.000
Saturated.Fat.cent	15.861	6.730	2.357	0.020
Trans.Fat.cent	0.495	1.074	0.461	0.646
Cholesterol.cent	-0.848	9.999	-0.085	0.933
Sodium.cent	-8.990	11.414	-0.788	0.432
Carbohydrates.cent	117.309	3.585	32.719	0.000
Dietary.Fiber.cent	2.998	1.097	2.734	0.007
Sugars.cent	-8.659	3.617	-2.394	0.018
Protein.cent	35.835	6.878	5.210	0.000
Vitamin.A....Daily.Value.cent	3.869	1.917	2.018	0.046
Vitamin.C....Daily.Value.cent	1.712	0.481	3.556	0.001
Calcium....Daily.Value.cent	6.326	2.771	2.283	0.024
Iron....Daily.Value.cent	3.436	2.202	1.560	0.121

The equation for our food dataset can be read as the following:

$Calories = 356.609 + 103.013(Total.Fat) + 15.861(Saturated.Fat) + 0.495(Trans.Fat) - 0.848(Cholesterol) - 8.990(Sodium) + 117.309(Carbohydrates) - 2.998(Dietary.Fiber) - 8.659(Sugars) + 35.835(Protein) + 3.869(Vitamin.A....Daily.Value) + 1.712(Vitamin.C....Daily.Value) + 6.326(Calcium....Daily.Value) - 3.456(Iron....Daily.Value)$

Because all of the variables have been standardized, they all have a mean of zero and a variance of one. This means that the coefficients for each model can be very easily compared with one another. Higher coefficients represent a greater relationship between the variable and the response, and lower coefficients represent a lesser relationship between the variable and the response.

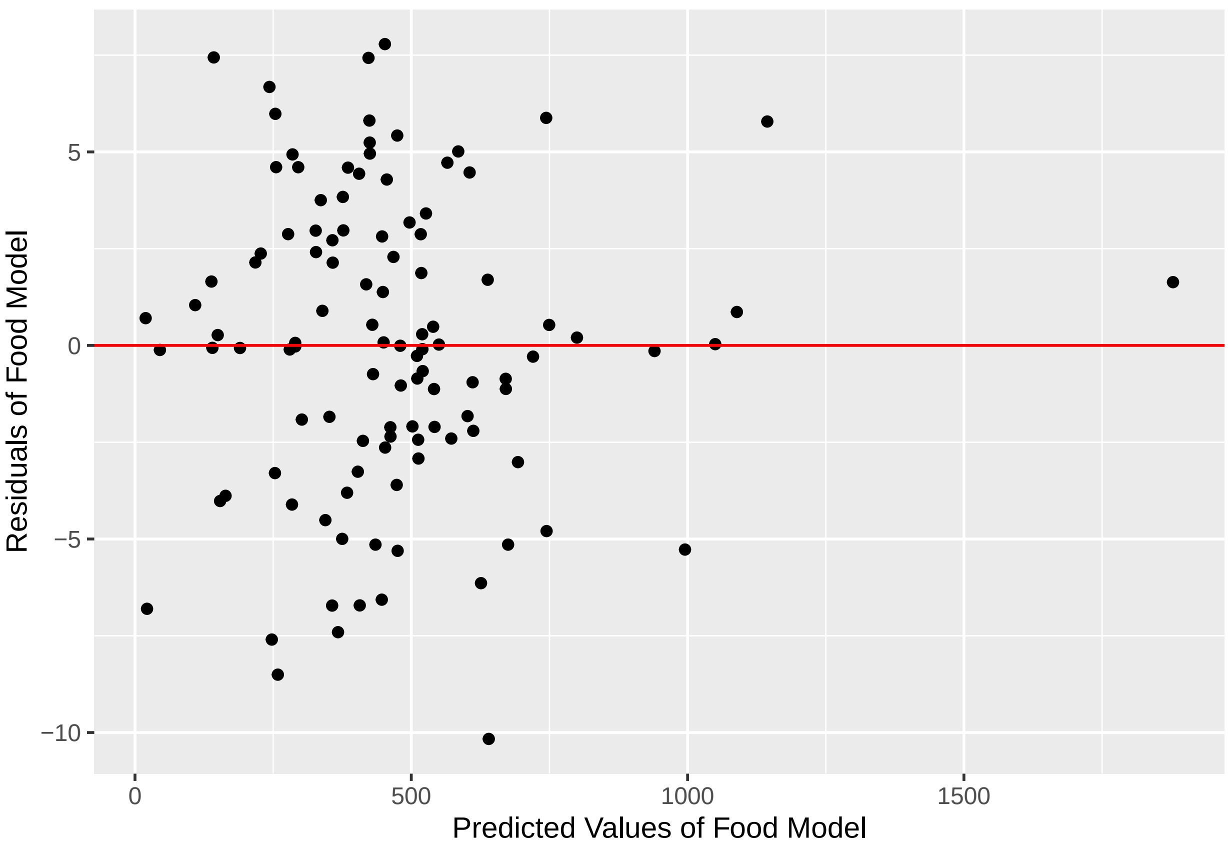
Undoubtedly, the estimators with the greatest relationship to calorie count in both foods and beverages are Total Fat, Carbohydrates, and Protein. As an aside, we did not hypothesize that calcium would have a significant effect on beverages. Thinking back, this result makes sense because milk-based beverages are both high in calcium and calories.

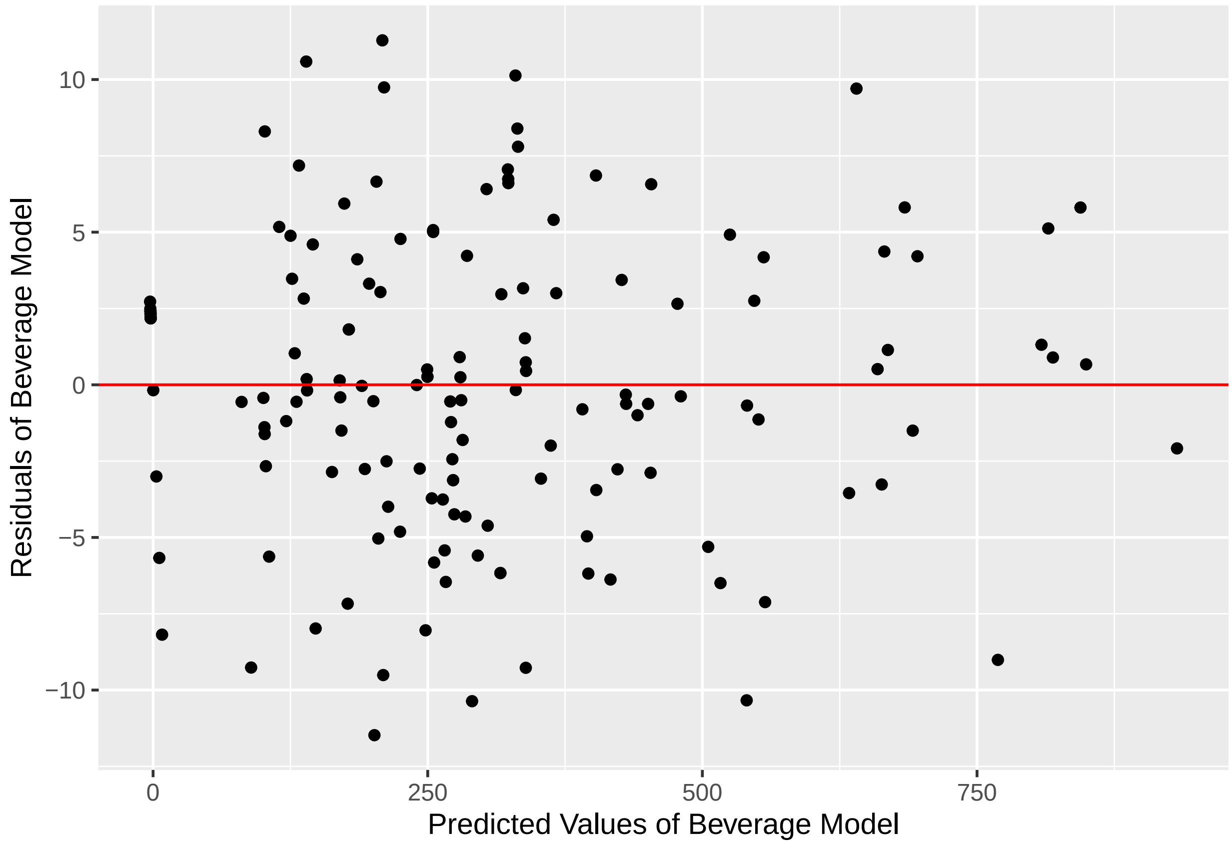
When first starting the analysis we decided to divide our data between food and drink items on the assumption that they have different caloric makeups. We split our dataset between food menu items (Breakfast, Beef & Pork, Chicken & Fish, Salads, Snacks & Sides, Desserts) and drink menu items (Coffee and Tea, Smoothies and Shakes, Beverages). For food menu items, we hypothesized total carbohydrates (grams) was the most accurate estimator of calories. For drink menu items, we hypothesized that total sugar (grams) was the most accurate estimator of calories. Upon completing our Exploratory Data Analysis and experiment, we concluded that best 3 estimators (nutritional attributes) were the same for food and beverage items. Thus, splitting the data did not lead to different results. If the same experiment were performed without splitting the dataset, we can expect Total Fat, Carbohydrates, and Protein to remain the highest coefficients.

Checking Assumptions

Before accepting our results as truth, we must inspect where the data and the model come from. There are certain assumptions we were accepting as truth when conducting our experiment. In order to confirm our results, we must verify that our prior assumptions are correct.

Constant Variance AISHA can you write about how these plots were created and explain that the constant variance assumption is satisfied

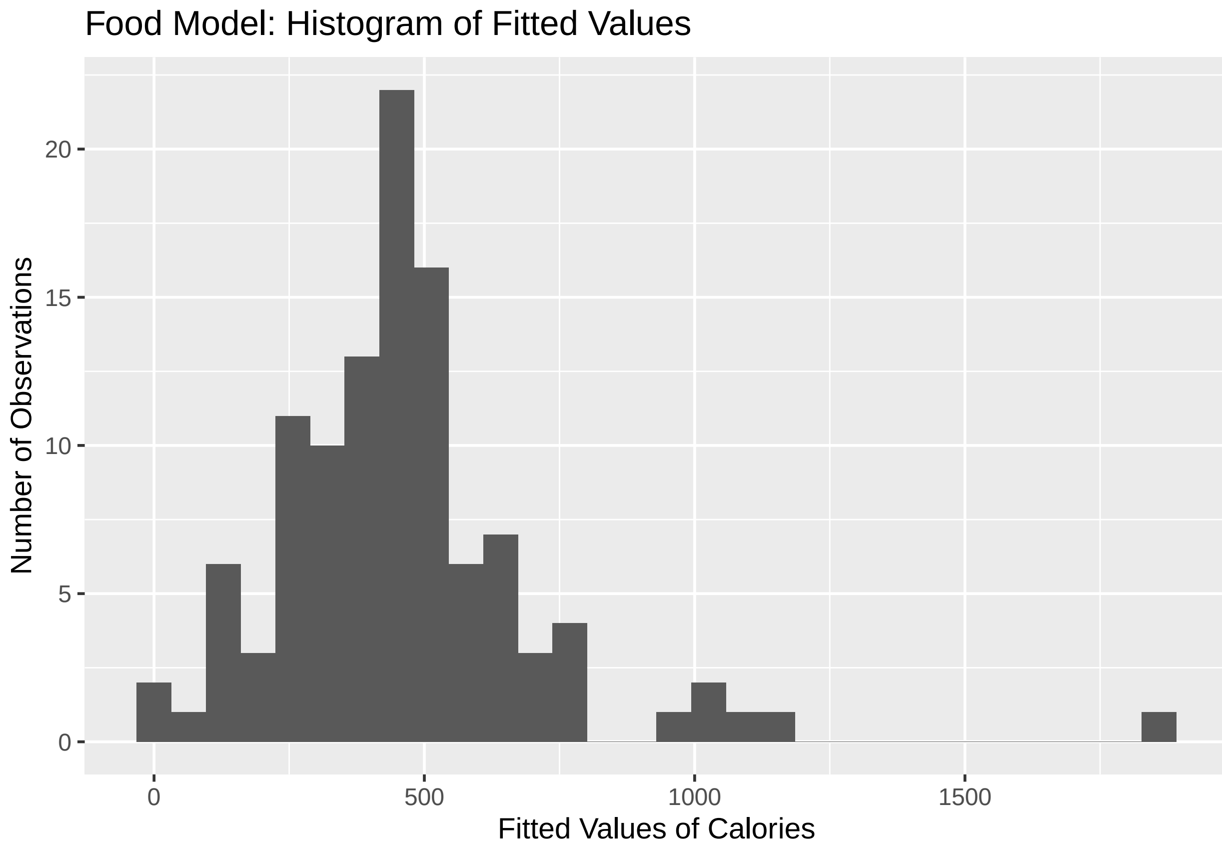


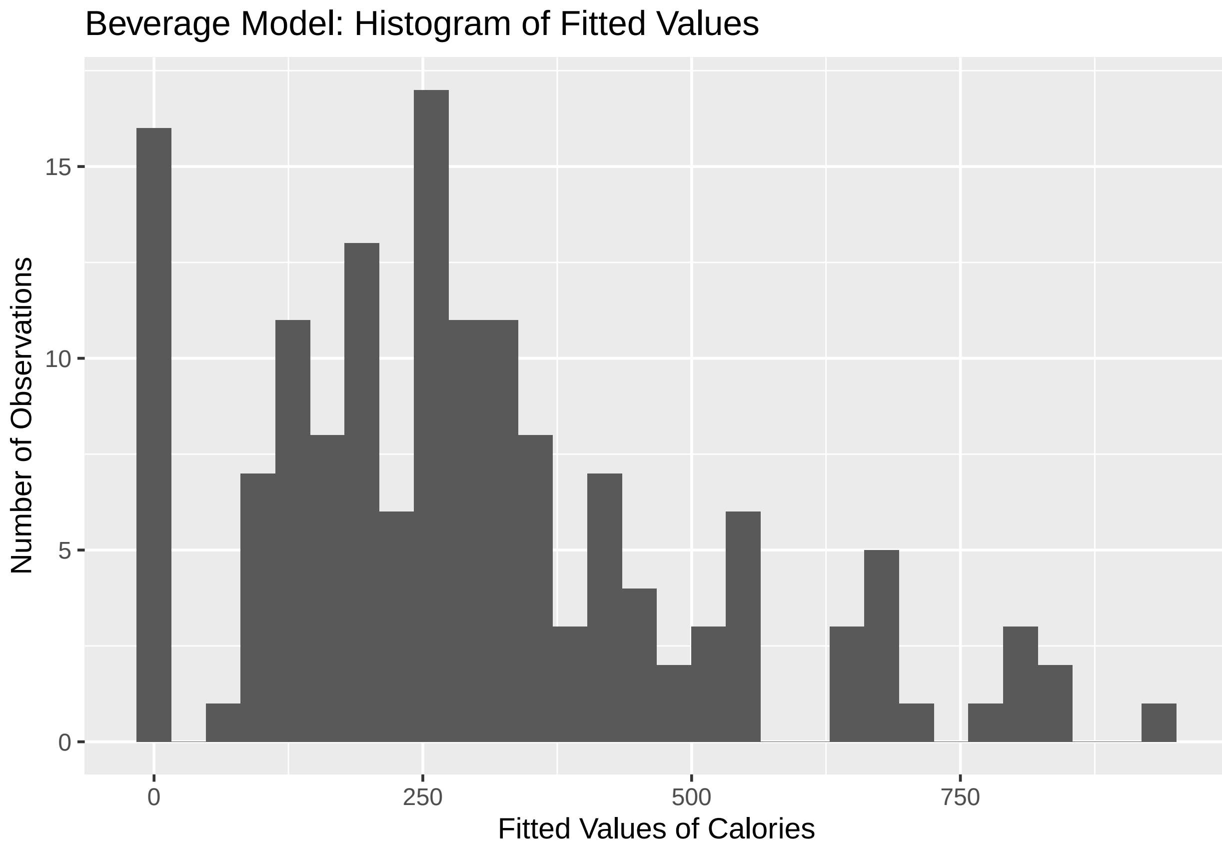


Linearity

The plots on page 4 and 5 show that for both food and beverages, there are linear relationships between calories and other nutritional attributes. This tells us that the linearity assumption is satisfied.

Normality AISHA can you explain the same for the normality assumption


Food Model: Histogram of Fitted Values
Number of Observations
0
5
10
15
20
0
500
1000
1500
Fitted Values of Calories



Independence

A crucial assumption of linear regression is the independence of observations. Looking at how our data was collected will indicate if the independence assumption is satisfied or not. Given that our dataset consists of nutritional attributes for each McDonalds menu item, each observation is independent. A menu item's observed nutritional attributes does not rely on other menu items. In part by FDA Menu Labeling Requirements (2020) the process by which our data was collected ensures data validity and that we are working with a random sample.