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 for 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

Exploratory Data Analysis

For this project, we understand that foods and beverages may have different predictors for their number of calories. Therefore, we will be splitting our dataset into two different dataframes: one for foods, and one for

beverages.

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.

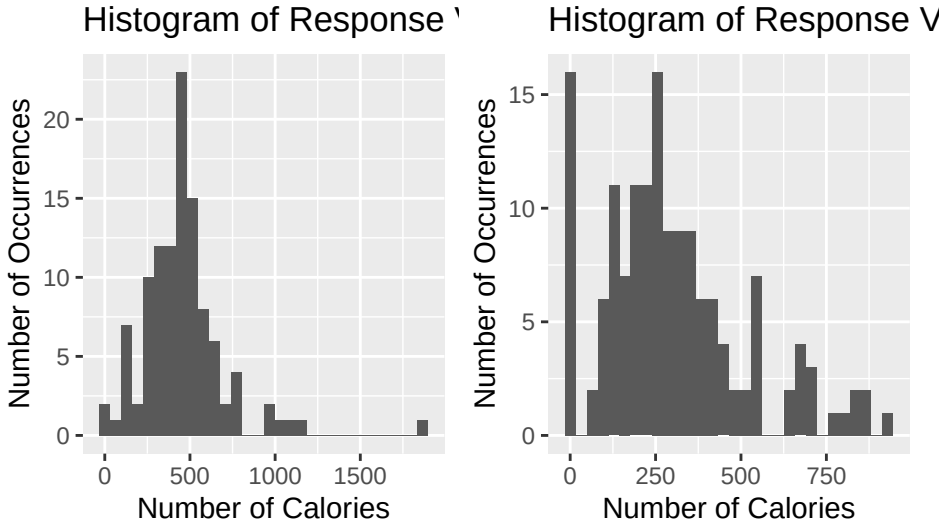
In order to accurately compare the estimators, we want to standardize them first. To do this, we make sure every estimator is centered around a mean of zero with a variance of one.

Next, we are dividing our categorical variables between food items (Breakfast, Beef & Pork, Chicken & Fish, Salads, Snacks & Sides, Desserts) and drink items (Coffee and Tea, Smoothies and Shakes, Beverages).

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.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



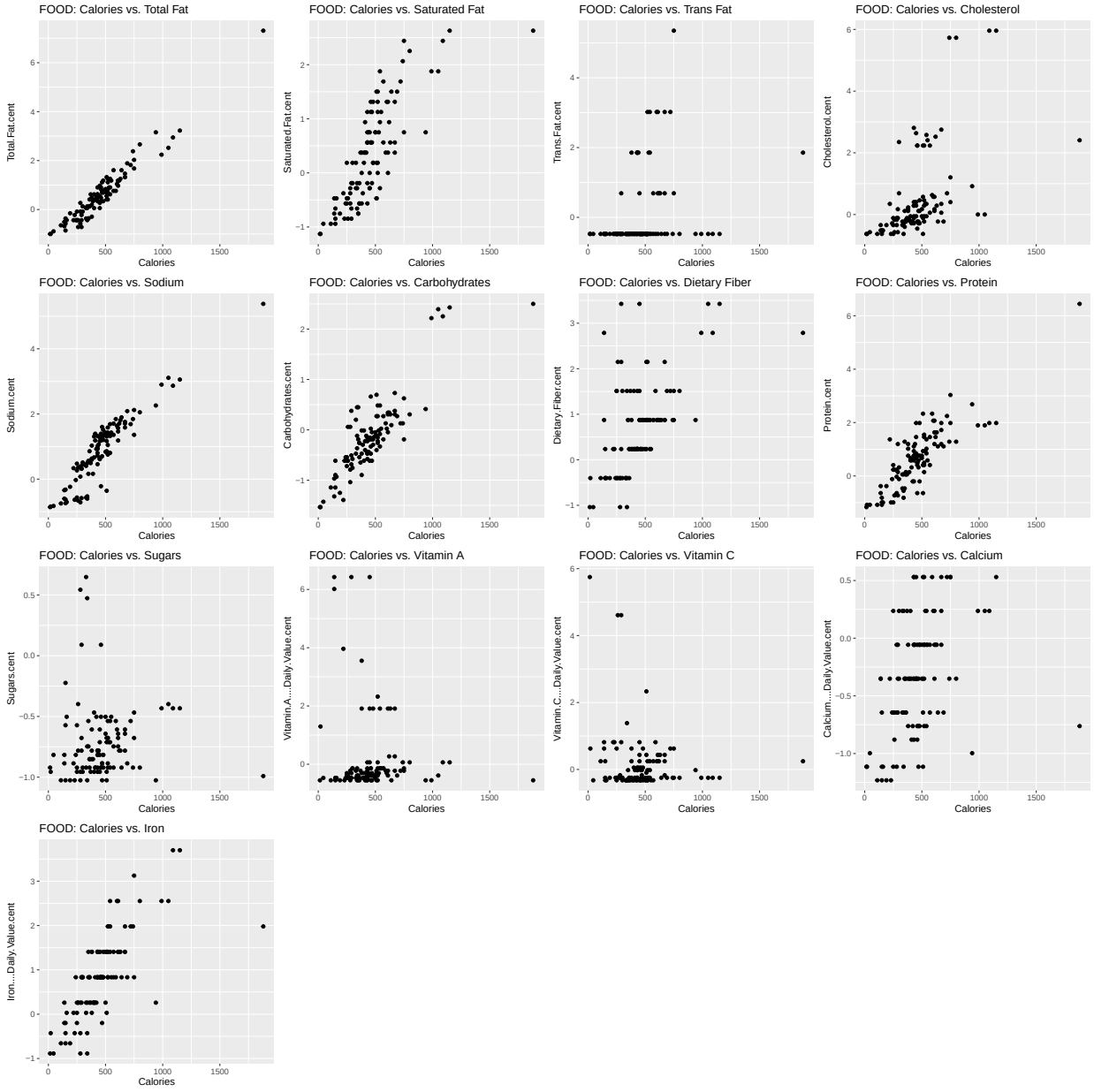
Now we calculate the appropriate summary statistics for calories (mean, median, standard deviation, IQR) for food items and drink items.

```
##       mean median  std_dev iqr
## 1 462.0909    445 249.3343 210
```

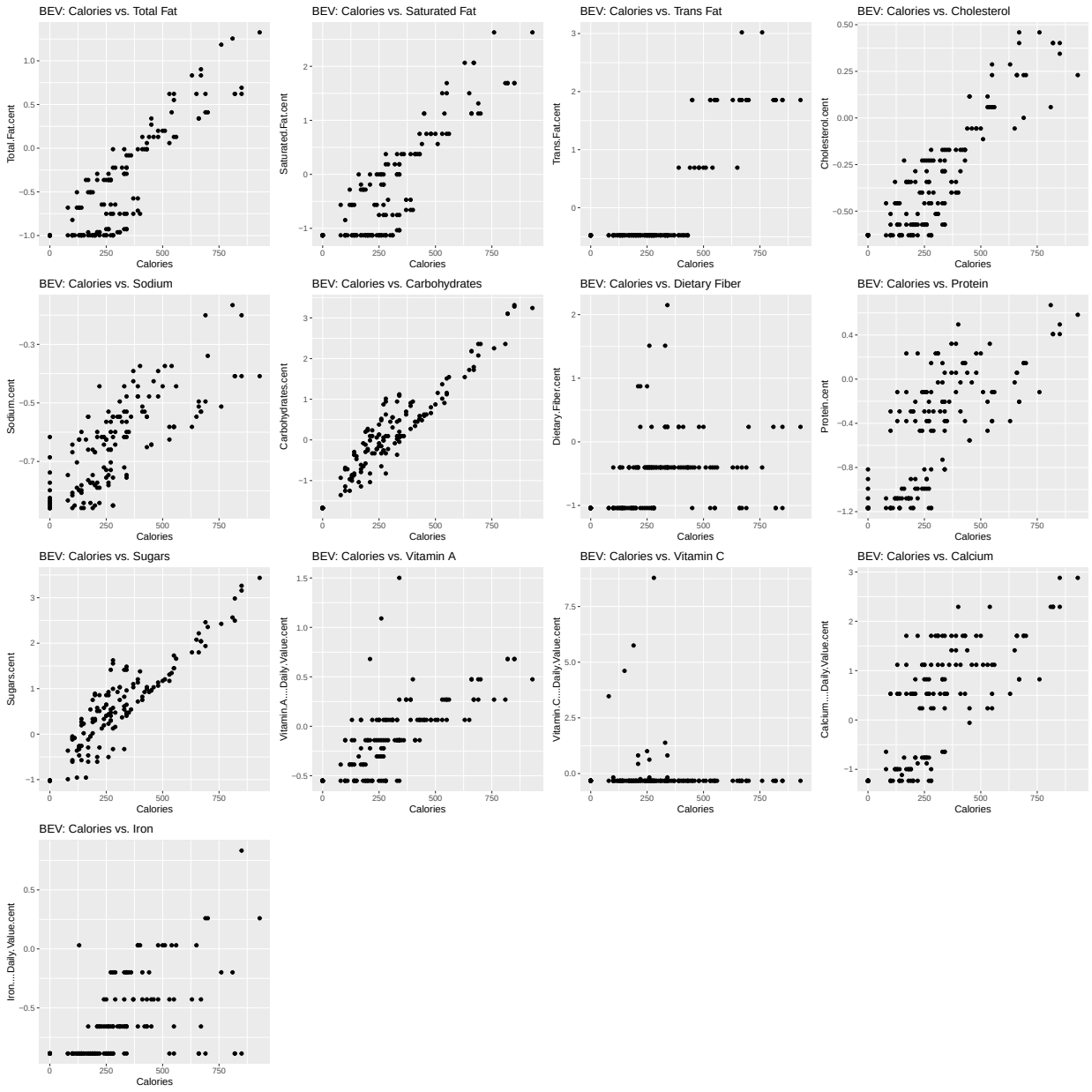
```
##       mean median  std_dev iqr
## 1 299.4667    270 208.8215 235
```

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.



Below is every plot that describes the beverage dataset.



Model Selection with AIC

The code below creates a linear model for McDonalds food menu items and displays the model output.

term	estimate	std.error	statistic	p.value
(Intercept)	-1.129	1.191	-0.949	0.345
Total.Fat	8.903	0.110	80.869	0.000
Saturated.Fat	0.645	0.275	2.344	0.021
Trans.Fat	1.311	1.529	0.858	0.393
Cholesterol	-0.009	0.005	-1.703	0.092
Sodium	-0.002	0.003	-0.605	0.547
Carbohydrates	4.128	0.072	57.453	0.000

term	estimate	std.error	statistic	p.value
Dietary.Fiber	-1.350	0.517	-2.614	0.010
Sugars	-0.060	0.098	-0.616	0.539
Protein	3.979	0.108	36.800	0.000
Vitamin.A....Daily.Value.	0.017	0.016	1.046	0.298
Vitamin.C....Daily.Value.	0.042	0.020	2.130	0.036
Calcium....Daily.Value.	-0.054	0.082	-0.657	0.512
Iron....Daily.Value.	-0.066	0.135	-0.490	0.625

The code below creates a linear model for McDonald's drink menu items and displays the model output.

term	estimate	std.error	statistic	p.value
(Intercept)	-1.074	0.946	-1.134	0.259
Total.Fat	9.049	0.082	110.387	0.000
Sodium	-0.052	0.017	-2.983	0.003
Carbohydrates	4.347	0.113	38.384	0.000
Sugars	-0.477	0.117	-4.069	0.000
Protein	3.794	0.566	6.703	0.000
Vitamin.A....Daily.Value.	0.158	0.079	2.001	0.047
Vitamin.C....Daily.Value.	0.044	0.018	2.452	0.015
Calcium....Daily.Value.	0.261	0.158	1.652	0.101
Iron....Daily.Value.	0.765	0.212	3.613	0.000

Now, we will select a model for food items using AIC. We are using the step function in R to conduct backward selection using AIC as the selection criterion, and storing the selected model as food_model_select_aic. Finally, we display the coefficients of the selected model.

For food products, the predictors that give us the best model for predicting calorie count are: -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.

Now, we will select a model for drink items using AIC. We are using the step function in R to conduct backward selection using AIC as the selection criterion, and storing the selected model as bev_model_select_aic. Finally, we display the coefficients of the selected model.

For Beverages, the predictors that give us the best model for predicting calorie count are: -Total.Fat -Sodium -Carbohydrates -Sugars -Protein -Vitamin.A....Daily.Value. -Vitamin.C....Daily.Value. -Calcium....Daily.Value. -Iron....Daily.Value.

The code and its output below show us that the best predictors of food calories in order are Total Fat, Carbohydrates, and Protein.

The code and its output below show us that the best predictors of beverage calories in order are Carbohydrates, Total Fat, and Protein

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 predictor of calories. For drink menu items, we hypothesized that total sugar (grams) was the most accurate predictor of calories. Upon completing Exploratory Data Analysis, Regression, and Model Selection, our final models (food_model_aic and bev_model_aic) indicated that best 3 predictors (nutritional attributes) were the same for food and beverage items. We then decided it would be best for our final model to encompass food and beverage data together. Our modeling objective benefits from a bigger dataset, as predictions are more accurate when working with a bigger dataset.

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.