Nutrition Hacks: Models for Fast Food Nutritional Data Analysis

Co Sou

INTRODUCTION:
This paper is an anlysis of nutrition facts collected by the MenuStat project, a database compiled by the NYC Department of Health and Mental Hygiene. It offers multi-year records of items sold by well-known national restaurant chains and their nutrition facts. The goals here are 1) to create models that can offer predictive insights in the nutritional breakdown of certain profiles and 2) to examine the changes in how the nutrition facts might change for an item throughout the years. The latter is achieved through a selected case study involving items sold by Burger King and its longitudinal effects. It is the hopes that these offered models can educate the public on the importance of carefully following diet guidlines recommended by the FDA and showcase how statistical analysis can generalize ideas on what we should eat through a scientific inquiry.

DATA:
The data offered by MenuStat contains records of items sold by national fast food chains and their breakdown of basic nutritional components. Variables recorded include Serving Size, Calories, Fat, Soidum, Carbs content and etc. What's different from this project to other databases of similar information is 1) its focus on fast food restuarants and 2) careful records of items throughout consecutive years. Hence, this database presumes a longitudinal design. Here, we analyze the contents of items for a 6 year period from 2012 to 2017 consisting of about 163,000 observations. Variables used in each model to follow will be displayed to ensure the reader know the exact variables used in its construction. For the longitudinal part of the paper, only items from Burger King will be used and the exact item names will be revealed as well.

ANALYSIS:
PART I: LOGISTIC REGRESSION
The simplest method for prediction to begin the analysis is that of logistic regression. The question asked here is whether the nutrition facts can determine whether an item is classified as a Kids Meal or not. Intuition tells us that Kids meals are smaller in size and would be designed for lesser consumption than that of regular items. A logistic regression model is employed to answer this question. Models of this kind requires complete information, and therefore all observations with missing variable values are removed. Additionally, the model is sensitive to extreme values that can end up obscuring the coefficients output and a choice of limiting the data to reasonable constraints is made. This gives us a usable dataset of about 76,000 observations. No transformation is done on the variables as we are not bound to normality assumption for logistic regression. Model selection is also performed and the result shows that the saturated model is preffered since it has a lower AIC than the models with reductions. Finally, a confusion matrix is computed to let us know how well the model performed.

This saturated model actually has the lower AIC (26364.64) compared to the alternatives.

```
## glm(formula = Kids_Meal ~ Food_Category + Serving_Size + Calories +
##     Carbohydrates + Protein, family = binomial, data = menu_total.subset2.na.1)
```

```
##                          (Intercept)           Food_CategoryBaked Goods
##                         -0.671259904                       -3.336604228
##              Food_CategoryBeverages               Food_CategoryBurgers
##                         -1.253589366                        1.259971576
##               Food_CategoryDesserts               Food_CategoryEntrees
##                         -1.337239222                        0.937306877
##         Food_CategoryFried Potatoes                 Food_CategoryPizza
##                          0.906008719                       -2.713565144
##                Food_CategorySalads            Food_CategorySandwiches
##                         -1.165928421                        0.184082264
```

```
##                   Food_CategorySoup Food_CategoryToppings & Ingredients
##                        -2.605378952                        -1.777386337
##                         Serving_Size                            Calories
##                        -0.003463029                        -0.005665337
##                        Carbohydrates                             Protein
##                         0.010068736                        -0.024775140
```

The model's confusion matrix is returned below. It has a success rate of about 95%, so in terms of performance it is a desirable model.

```
##          predout
## Kids_Meal    no   yes
##       No  72152     5
##       Yes  3778     1
```

We also offer an alternative paradigm of splitting the data into training and testing sets. With over 75,000 observations, this data is somewhat large and their is a chance of the model overfitting on error. The following is the same model as above done under the context of a validation set approach. It returns a model of AIC (652.6973), which is much less than the regular logistic model without training. The confusion matrix is also given, which has a success rate of about 94%, similar to the regular logistic regression. In both models, coefficients returned are similar to each other. In conclusion, we would prefer the trained logistic regression under validation over the regular logistic regression.

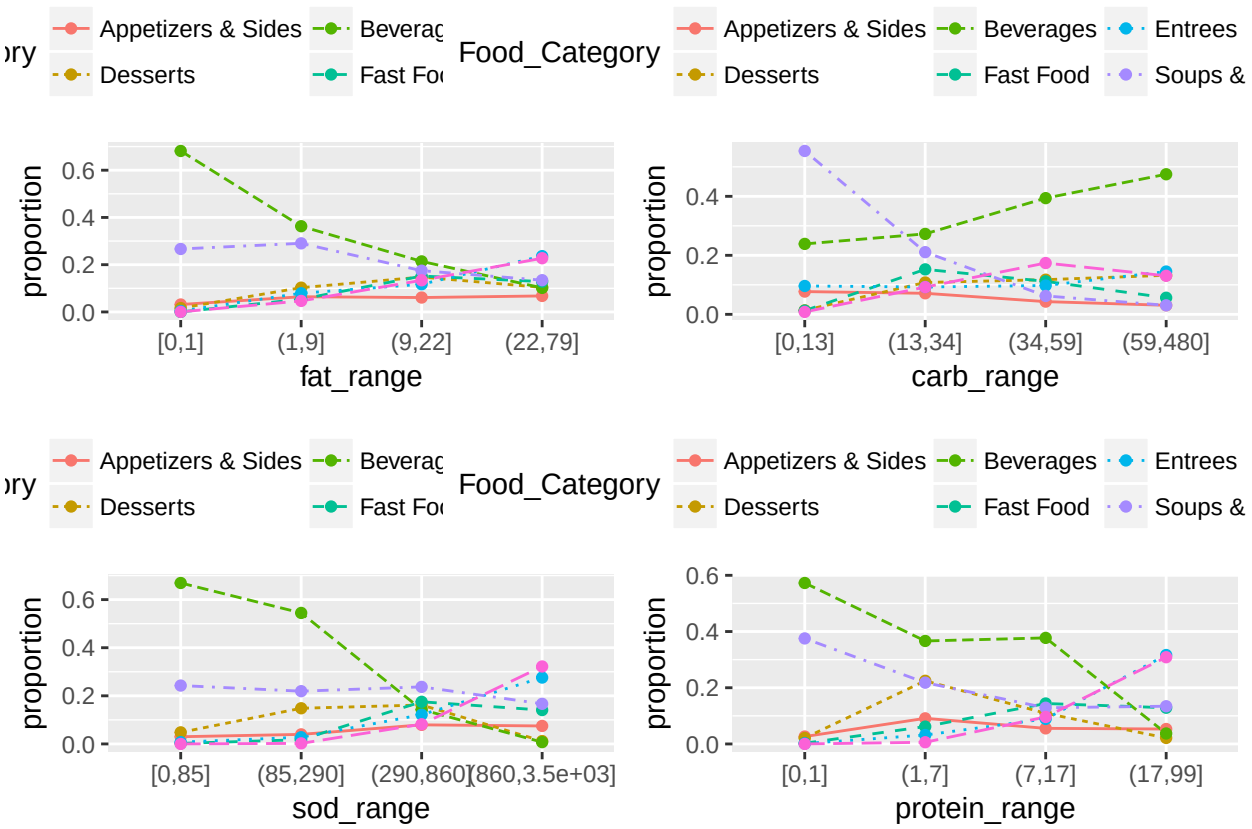
```
## glm(formula = Kids_Meal ~ Food_Category + Serving_Size + Calories +
##     Total_Fat + Carbohydrates + Sodium + Protein, family = binomial,
##     data = menu_total.subset2.na.1, subset = train.logistic)
```

```
##                           (Intercept)            Food_CategoryBaked Goods
##                         -3.044604e+00                         -1.573645e+01
##               Food_CategoryBeverages                Food_CategoryBurgers
##                         -2.156345e+00                          3.019033e+00
##                Food_CategoryDesserts                Food_CategoryEntrees
##                         -8.151495e-01                          3.461843e+00
##         Food_CategoryFried Potatoes                   Food_CategoryPizza
##                         -1.620951e+01                          2.509975e+00
##                  Food_CategorySalads             Food_CategorySandwiches
##                         -1.206125e+01                          2.599266e+00
##                    Food_CategorySoup Food_CategoryToppings & Ingredients
##                         -1.443362e+01                         -1.401814e+01
##                          Serving_Size                             Calories
##                         -1.810808e-03                          4.029089e-03
##                             Total_Fat                           Carbohydrates
##                         -4.210604e-02                         -1.258920e-02
##                                Sodium                                Protein
##                         -1.671818e-04                         -1.339878e-01
```

```
##                Kids_Status
## glm.log.pred        No   Yes
##   Kids Meal:Yes 62347  3707
##   Kids Mean:No     35     3
```

PART II: MULTINOMIAL REGRESSION

As a generalization of the logistic regression model to many factor level response, the multinomial model allows us to ask what relationships exist between the nutrition facts and their corresponding item's food classification. Here, the multinomial model is created for the Food Category variable as a reponse. A choice is made to reduce the number of factor levels into 7 catgeories from 12 categories for simplicity's sake. As before, all missing and extreme values are omitted as a condition for running this model.



First we give a visulization of the relationship between the proportions of an item type relative to the ranged intervals cross tabulated to each nutrition type. Some important conclusions include 1) for Beverages there is a an increase in proportion as carbohydrates range incease; unsurprising since most of the drinks sold must have high level of added sugar 2) Soups and Salads tend to decrease in carbs and protein 3) Sandwiches increase in sodium steadily.

The goal here is to treat the food category as the response and see if the other nutritional variables can predict it. The saturated multinomial model is reduced but it turns out that the saturated model has the lowest AIC (303049.1).

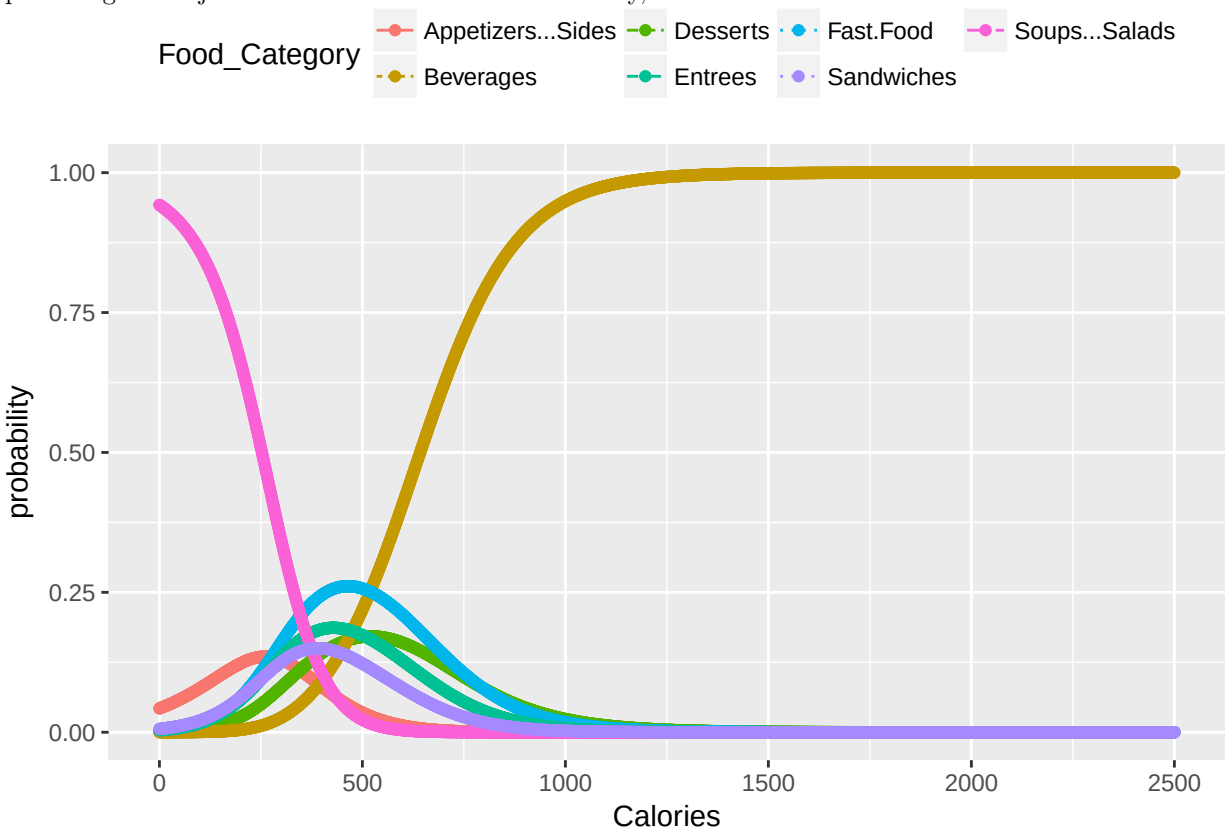
```
## multinom(formula = Food_Category ~ Calories + Total_Fat + Sodium +
##     Carbohydrates + Protein, data = menu_total.subset1.na.2)

##                (Intercept)     Calories   Total_Fat        Sodium
## Desserts         0.5771941  0.010289857 -0.07917339 -0.0043901337
## Beverages        2.1369235  0.017177306 -0.23999332 -0.0102253688
## Fast Food       -0.3092259  0.008896974 -0.07938100 -0.0010313311
## Entrees         -0.9773142  0.007842137 -0.09409458 -0.0006421163
## Soups & Salads   2.3644063 -0.007083403  0.04879070 -0.0002154239
## Sandwiches      -1.0602936  0.006443093 -0.09248194  0.0002059824
##                Carbohydrates     Protein
## Desserts          0.019121718 -0.09877584
## Beverages         0.020284050 -0.02174248
## Fast Food        -0.015149224  0.02046179
## Entrees          -0.011189138  0.07171476
## Soups & Salads   -0.005966419  0.03053321
## Sandwiches       -0.006952904  0.05154987
```

We interpret the coefficients as following: The baseline is the Appetizers catgeory. If calories increase by one unit, your chances of staying in the appetizers category are higher compared to staying in the Soup and

Salads.In english this means as calories increase by a level, there is a greater chance of appetizers increasing its calorie count than for the Salads category.Likewise, if carbs increase by one unit there is a greater chance that both beverages and desserts would increase in carbs count than for the Appetizers group. Another conclusion: If fat increase by one unit, there is a greater chance that salads will increase its fat count than for the appetizers. This last result seems counterintuitive. After all, shouldn't it be the other way around? Perhaps the salads sold in fast food chains aren't as healthy as we think as they are loaded with fried meats, cheese, and heavy sauces.

Of interest is the following probability plot. We see that Soups and Salads tends to decrease in Calories and are mostly in the 0-500 calorie range. On the other hand, Beverages tends to increase in Calories, some are in the mid-hundreds range, which is eye-opening. Most of the food types fall in the moderate 300-700 calorie range. Notice the tapering of the right ends. This is because most of the items sold by these chain restaurants are appropriately proportioned for a single consumption. Recall that the Daily Value recommendation is 2000 Calories a day. If someone eats a single item of that much Calories, then they would have used all of the daily percentage limit just for one meal item. But thankfully, that isn't the case here since the items were filtered.



The confusion matrix for the model is shown below. It has a success rate of about 55%. Depending on the context, we might want something more accurate since biostatistical/clinical work is of higher stake.

```
##                       menu_total.subset1.na.2$Food_Category
## predict(mmod.reduced) Appetizers & Sides Desserts Beverages Fast Food
##    Appetizers & Sides                  90        2         0        82
##    Desserts                           259     3533      1635       495
##    Beverages                         1211     7289     38319       304
##    Fast Food                          336       56        14       592
##    Entrees                            493       75        19      2568
##    Soups & Salads                    4146     1190      6034      5942
##    Sandwiches                         937      221        37      1177
##                       menu_total.subset1.na.2$Food_Category
## predict(mmod.reduced) Entrees Soups & Salads Sandwiches
```

```
##    Appetizers & Sides      44             15         11
##    Desserts               458            149         36
##    Beverages              884           5404         71
##    Fast Food              894            195        854
##    Entrees               5106           1641       3679
##    Soups & Salads        3723          20668       3453
##    Sandwiches            3293            963       5406
```

To avoid the possibility of overfitting, the above multinomial model is also done under a validation set approach. It yields a confusion matrix of similar success rate (54.3%). But this model has a lower AIC of 256346.1. We prefer the trained multinomial model under validation as our final model.

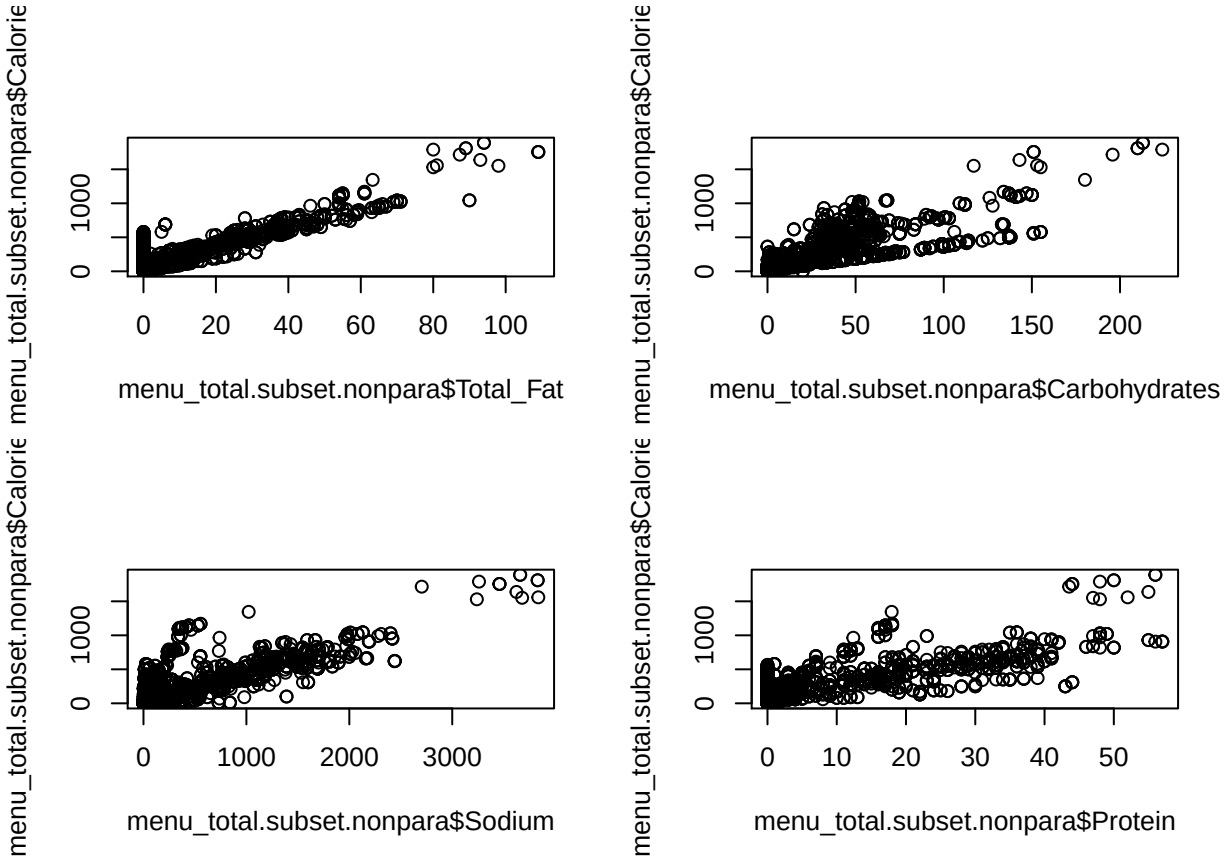
```
## multinom(formula = Food_Category ~ Calories + Total_Fat + Sodium +
##     Carbohydrates + Protein, data = train.multimod)
```

```
##
## pred.mmod2.class     Appetizers & Sides Desserts Beverages Fast Food
##   Appetizers & Sides                 45        1          0        19
##   Desserts                           41      659        344        97
##   Beverages                         244     1495       7715        79
##   Fast Food                          55        4          2       108
##   Entrees                           225        9          4       635
##   Soups & Salads                    827      230       1198      1248
##   Sandwiches                        155       60          5       170
##
## pred.mmod2.class     Entrees Soups & Salads Sandwiches
##   Appetizers & Sides      41             24         20
##   Desserts               100             29          3
##   Beverages              182           1090         13
##   Fast Food              132             39         86
##   Entrees               1368            372       1125
##   Soups & Salads         808           4232        759
##   Sandwiches             514            117        890
```

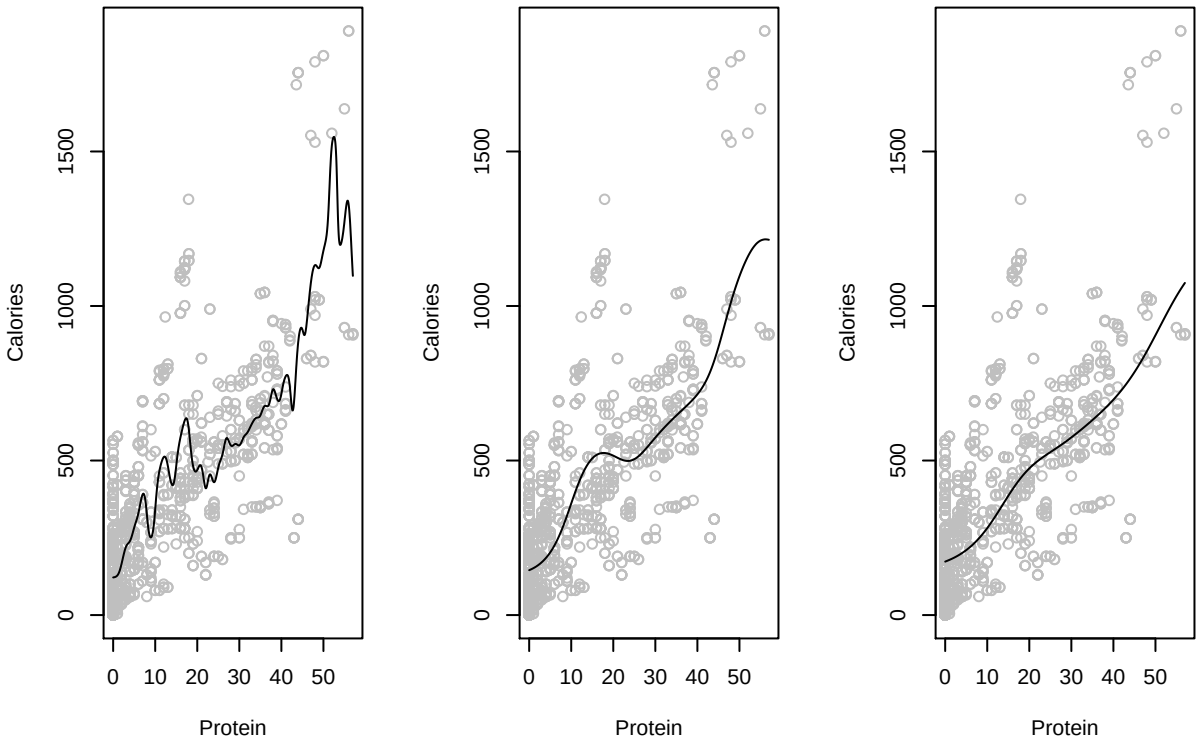
PART III: NONPARAMETRIC REGRESSION

When modelling univariate responses with only a single predictor, one has to be skeptical of whether just one predictor can truly be a strong predictor. But sometimes, that's all we have and sometimes that's what we are interested in. For approaches like this, non-parametric methods provide us a simple starting point for when we encounter unfamiliar places and don't want to assume anything about the underlying distribution of the data.

Below is a panel view of four plots showing the relationship between Calories as the response with the four major nutrition types as single predictors. For both Calories ~ Fat and Calories ~ Sodium, there seems to be a clear linear relationship albeit with some outliers in the tailed end. For Calories ~ Carbohydrates there is too much clustering at the left side of the plot; clustering algorithms in unsupervised learning might be helpful in analyzing this one (beyond the scope of this project). The plot showing Calories~ Protein shows the most promise. It's not clear if there is a linear relationship and points are more evenly spaced from each other. Nonparametric statistics can be employed to understand the dynamics of this relationship.

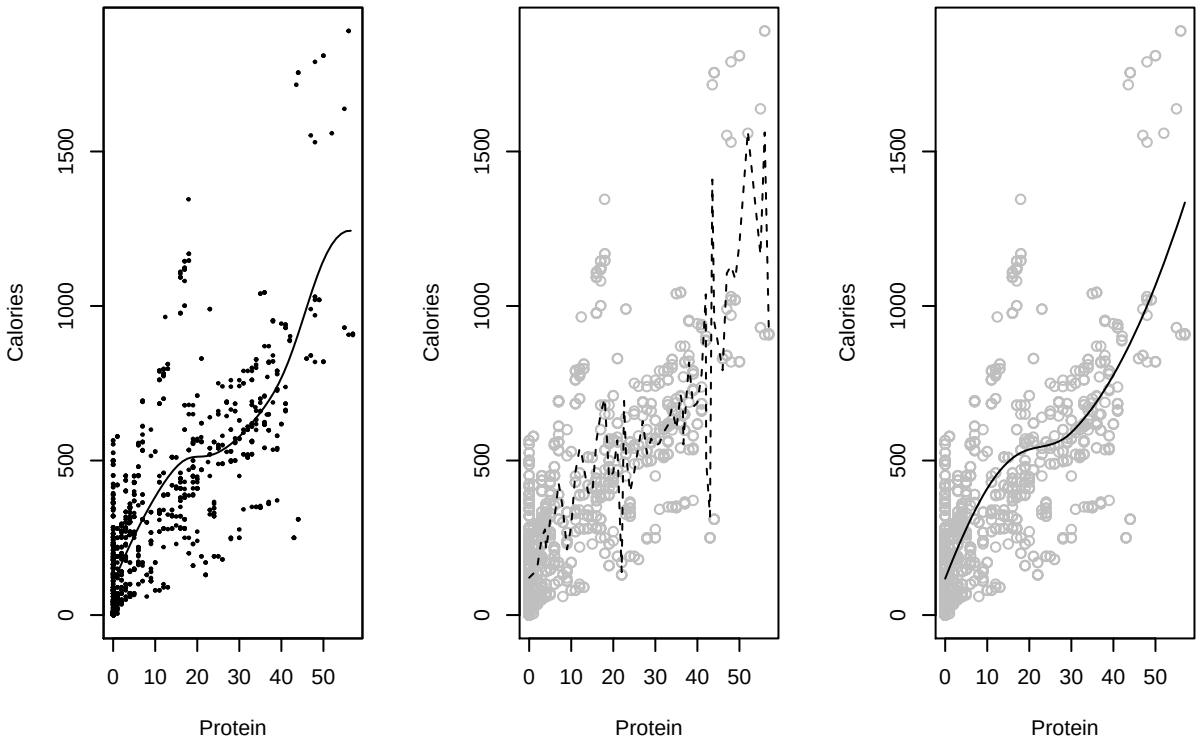


We begin with kernel estimation and the selection of bandwidth. A selection of three bandwidth parameters and their plots are shown. In the left most panel, the curve fitted is too varying and jagged but it does manage to reach some of the outliers. The right most panel shows a curve that is simply too linear; it doesn't seem to capture the variance of the points it passes through. The best is the middle panel where the curve is smooth enough and curvey enough to capture the variance. As for outliers, perhaps that is a choice to remove some of them before the analysis. Nevertheless, we prefer the middle bandwidth.

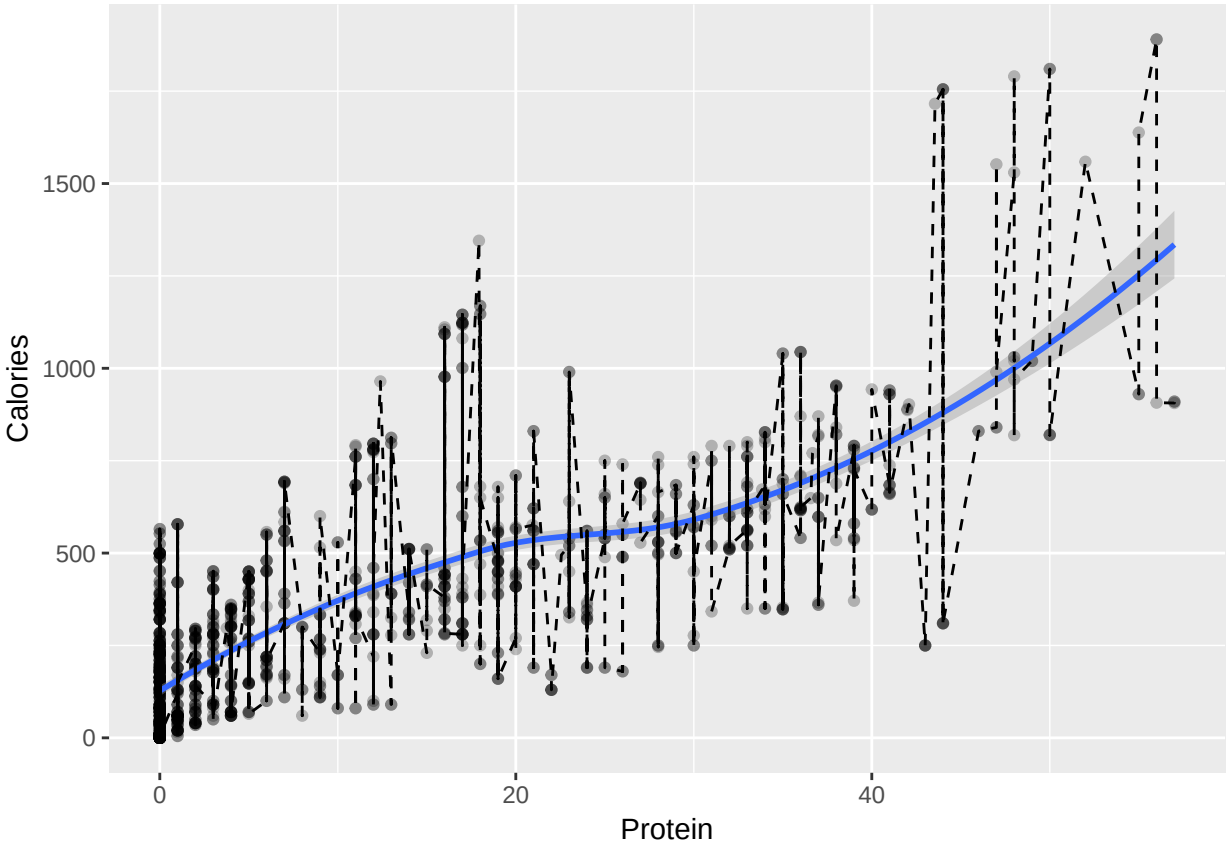


Bandwidth selection can also be estimated through optimization by cross validation. R performs the selection for us instead of experimenting with different bandwidths like we did above. Splines can also be used as well. Finally, we include Loess method.

The results show that both the cross validated and Loess methods return similar plots. The splines method gives a difficult to interpret and jagged curve that resemebles an overfitting behavior. The decision here is to use the Loess method based on local polynomials. It doesn't require that we specify a function in order to fit unlike the cross validation approach that does. We want to make as little assumption here and rely on as little dependencies as much as possible.

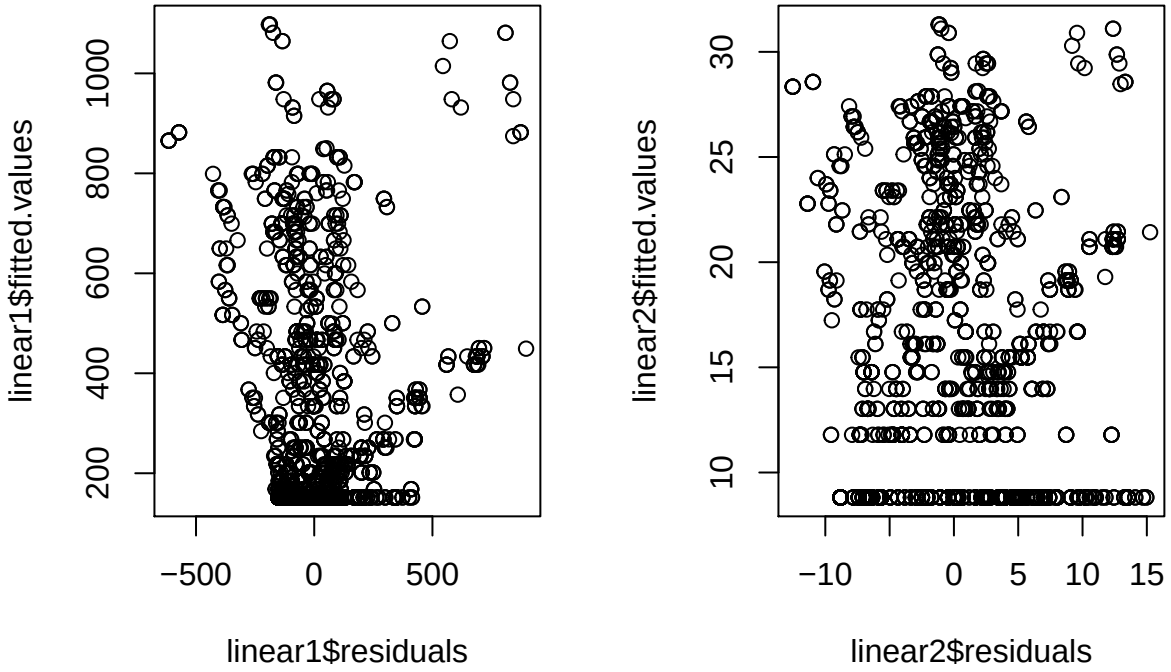


Confidence intervals can be created for the loess method. The graph below shows that confidence band isn't too far apart from the curve itsel, which is desirable.



The next step would be to try to get an analytic form of an equation that this curve represents. That is, of course, beyond the scope of this project and concludes the non parametric portion.

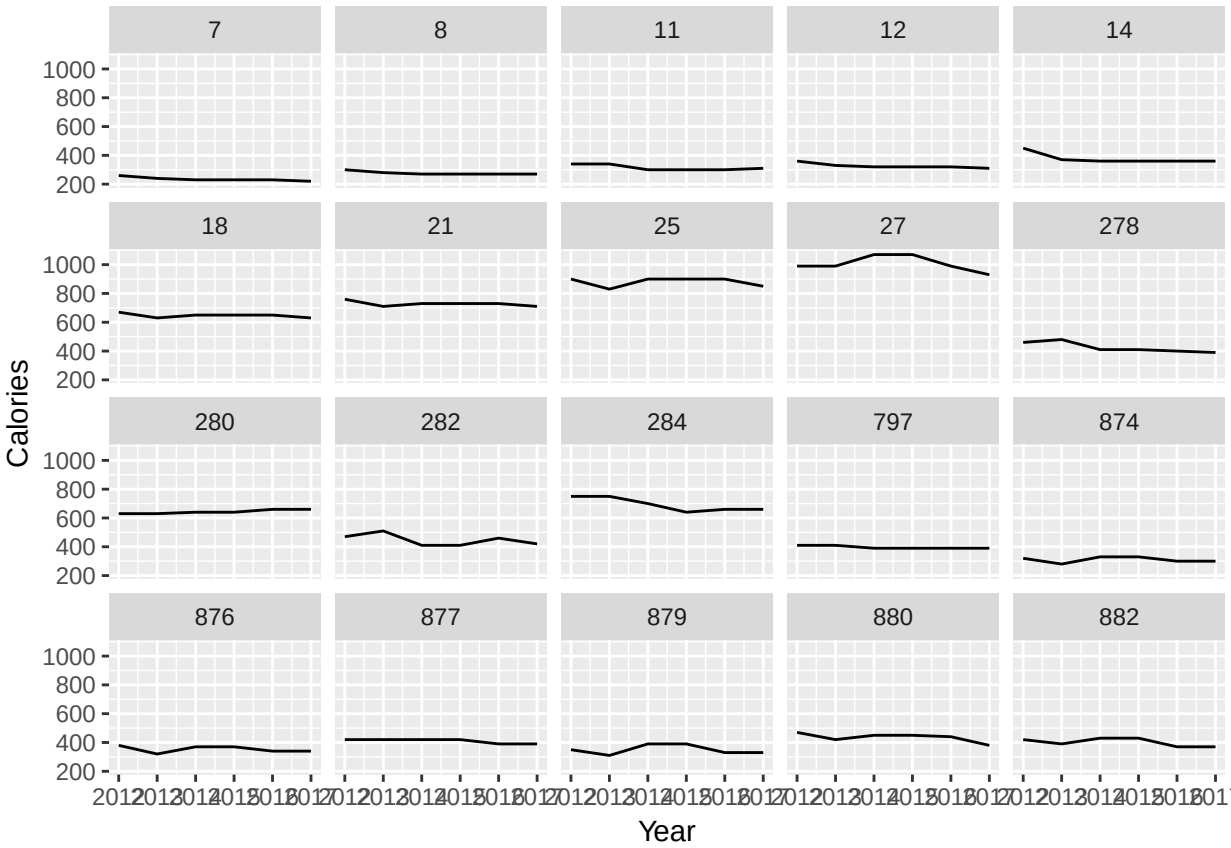
So how does this benefit our understanding of the problem? Compared to the two models shown below of Calories ~ Protein, one on the left panel is not transformed while the one on the right panel is square root transformed,at least we don't have to deal with problems on the normality assumption and worry about any nonstandard residual plot behaviors. To be fair though the square root transformed model is not the worst thing in the world, though the bottoming clusters serve a problem.



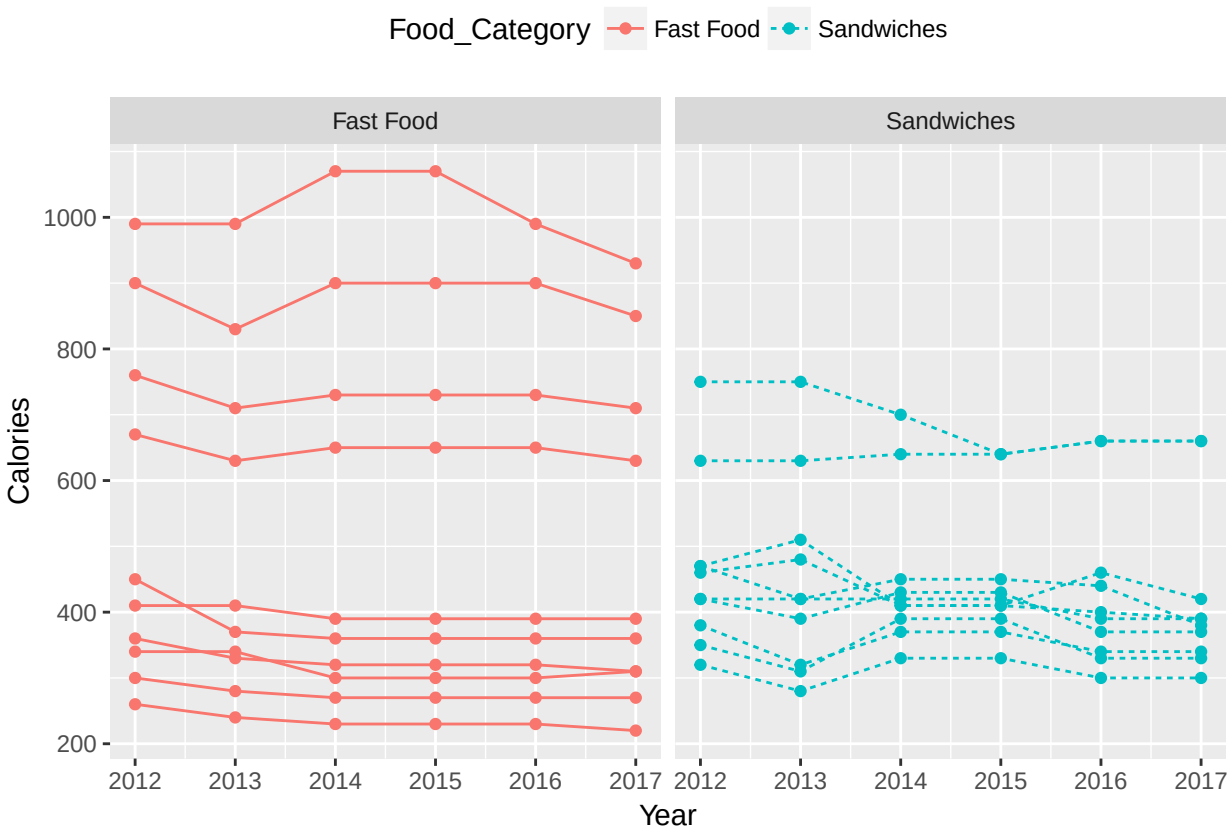
PART IV: LONGITUDINAL STUDY

The final part of this project explores the multi-year recording of selected items as a longitudinal design. Through a convoluted process the original data needs to be completely altered into an appropriate panel format. Afterward, some explorary graphics are produced. The most interesting part is a case study involving certain items from Burger King.

Below is a panel of the first twenty items offered by Burger King, as a subset of the entire 2012 to 2017 dataset. We are interested in how the Calories of these items change over the years. Most of the items don't seem to change calorie count by much. Indeed, items 7 (Hamburger), 8 (Cheeseburger), 280 (Original Chicken Sandwich w/ Mayo), and 797 (BK Veggie Burger) barely change at all. Other items have noticeable changes in calories. For example, items 27 (Double Whopper Sandwich w/ Cheese), 282 (Tendergrill Chicken Sandwich), and 284 (Tendercrisp Chicken Sandwich) show changes. Showing a panel of all the items would be impractical, of course.

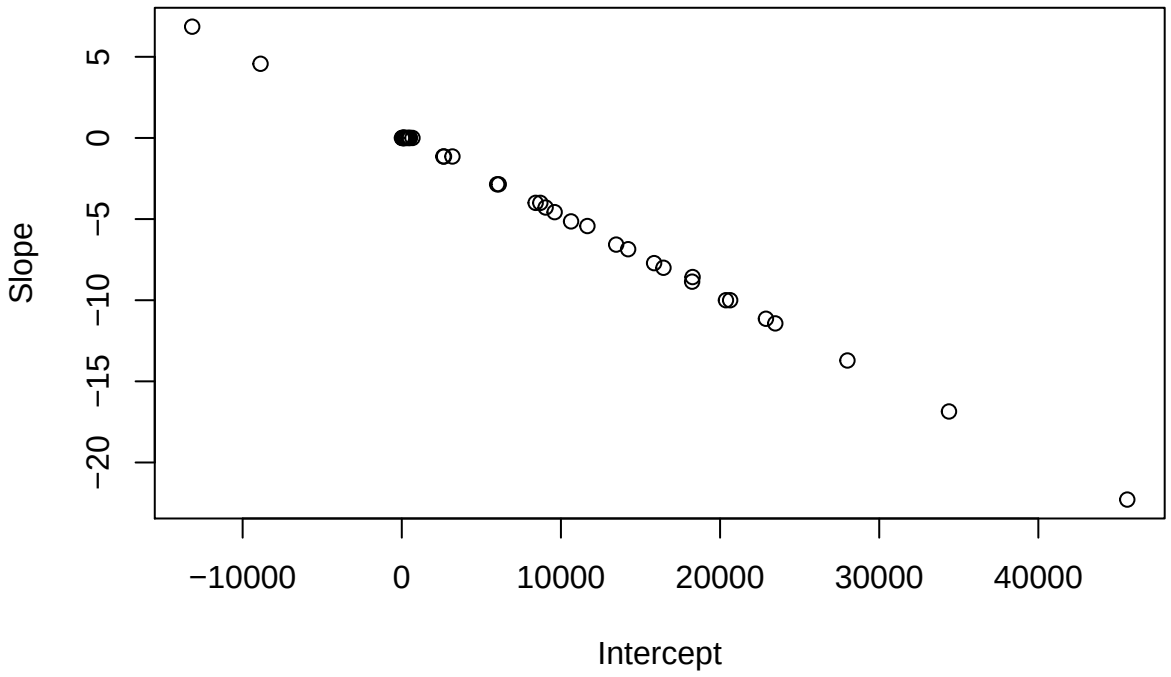


We can also separate the data into food types and see if the changes in calorie count has anything to do with their classfication. It's hard to tell just by looking at this but for the Sandwiches, there seems to be more variance when it comes to changes in calorie count by the intersecting lines.



In the modelling part of this section, the first 20 items has now been increased to the first 40 items of the Burger King subset along with their longitudinal components. The R library "lme4"" is required to perform this analysis. The reader may have trouble installing and attaching this package; in that case an installation directly from R CRAN repositories and restarting R is recommended.

A surprising result is given below. A linear regression is performed for each of the 40 items studied and their slopes and intercepts are plotted. Thee is a clear inverse relationship between the two. This suggests strong negative correlation between Calories and the change in calories over the year range. We cannot perform linear models on each of these qualities separately. We interpret this result as follows: Higher calorie items (intercept terms) will lead to a more rapid decrease in slope. This means higher caloried items tend to not change as much in calorie count. (Sort of like a ceiling effect)



At this point the analysis is concluded since this requires specialized knowledge of mixed effects modelling, which the R package lme4 is designed for. What's to gain from this preliminary analysis is that higher calorie items tend not to change much throughout the years and that longitudinal modelling allows us to incorporate time related predictors as an essential part of the data analysis process, whenever the data allows it.

CONCLUSION:
Both logistic regression and multinomial regression performed well for the prediction of categorical variables. Since there is a moderately large number of observations, both validation set versions of these models are preferred. Nonparametric curve estimation was used to deal with univariate problems with a single predictor without resorting to strong distribution assumptions and we concluded that the LOESS method is best. Finally, the longitudinal design of the data was taken into account and we discovered that there is a ceiling effect on high calorie items and their tendencies to not change as much.
The underlying philosophy of this project is to discover effects of nutritional facts of fast food sold with a focus on calorie counting. We've only scratched the surface of this topic and there are plenty of researchers in public health who work on these types of problem everyday, doing their best to ultimately influence the improvement in our nation's ever changing diets. Based on what've seen there are plenty of less than desirable qualities of some of the items showcased here. But we at least with the open source data available on this topic, the use of statistical procedures can clarify our understanding on deciding what to eat.

SOURCE: http://menustat.org/#/home