USDA Nutrition Data Analysis

Ruchi Neema, Snehal Vartak, Xiaotian Han, Zitong Li

November 27, 2017

Contents

Introduction

Most of the people today are conscious about the nutritional value of the foods they eat. USDA provides a nutrition data set for 8,789 different food items. The nutritient values are given for 100g of food serving. They also provide a separate dataset which classifies each of these food items in 25 different categories. We have extracted these food categories from the USDA website and included them as a new column in our dataset.

Data Description

1. Food_group - 25 different categories (eg. Baked Products, Dairy and Egg Products, etc)
2. Calories (in Kcal) - Continuous variable
3. Protein(g) - Continuous variable
4. TotalFat_(g) - Continuous variable
5. Carbohydrt_(g) - Continuous variable
6. Sodium_(mg) - Continuous variable
7. SaturatedFat(g) - Continuous variable
8. Cholestrl_(mg) - Continuous variable
9. Sugar(g) - Continuous variable

```
##   ï..NDB_No              Shrt_Desc             Food_group
## 1      1001       BUTTER,WITH SALT Dairy and Egg Products
## 2      1002 BUTTER,WHIPPED,W/ SALT Dairy and Egg Products
## 3      1003   BUTTER OIL,ANHYDROUS Dairy and Egg Products
## 4      1004            CHEESE,BLUE Dairy and Egg Products
## 5      1005           CHEESE,BRICK Dairy and Egg Products
## 6      1006            CHEESE,BRIE Dairy and Egg Products
##   Calories.in.Kcal. Protein.g. TotalFat_.g. Carbohydrt_.g. Sodium_.mg.
## 1               717       0.85        81.11           0.06         643
## 2               718       0.49        78.30           2.87         583
## 3               876       0.28        99.48           0.00           2
## 4               353      21.40        28.74           2.34        1146
## 5               371      23.24        29.68           2.79         560
## 6               334      20.75        27.68           0.45         629
##   SaturatedFat.g. Cholestrl_.mg. Sugar.g. Calcium.mg. Iron_.mg.
## 1          51.368            215     0.06          24      0.02
## 2          45.390            225     0.06          23      0.05
## 3          61.924            256     0.00           4      0.00
## 4          18.669             75     0.50         528      0.31
## 5          18.764             94     0.51         674      0.43
## 6          17.410            100     0.45         184      0.50
##   Potassium_.mg. VitaminC.mg. VitaminE.mg. VitaminD.microg.
## 1             24            0         2.32              0.0
## 2             41            0         1.37              0.0
## 3              5            0         2.80              0.0
## 4            256            0         0.25              0.5
## 5            136            0         0.26              0.5
## 6            152            0         0.24              0.5
```

Research Questions

Regression Problem

If we look at the nutrition label on any of the food items we buy, we can see that under the total calories it gives each nutrient source that contributes to the calorie value for that food. Based on this understanding, we built a model to see if the nutrients in our data set can succesfully explain the total calories for all foods. From this model, we will know the number of nutrient tests required to get the approximate calories in any new food item.

Classification problem

As we have added the food groups to the original USDA nutrient data, we can pose this as a classification problem to identify the food group given the nutrient content of that food item.

Exploring the dataset

Looking at the data set summary

```
##    ï..NDB_No      Shrt_Desc
##  Min.   : 1001   Length:8790
##  1st Qu.: 9086   Class :character
##  Median :14428   Mode  :character
##  Mean   :15664
##  3rd Qu.:20143
##  Max.   :93600
##
##                                Food_group   Calories.in.Kcal.
##  Beef Products                    : 967   Min.   :  0.0  
##  Baked Products                   : 879   1st Qu.: 91.0  
##  Vegetables and Vegetable Products: 836   Median :191.0  
##  Soups, Sauces, and Gravies       : 465   Mean   :226.3  
##  Lamb, Veal, and Game Products    : 464   3rd Qu.:337.0  
##  Sweets                           : 463   Max.   :902.0  
##  (Other)                          :4716                  
##    Protein.g.     TotalFat_.g.    Carbohydrt_.g.    Sodium_.mg.   
##  Min.   : 0.00   Min.   :  0.00   Min.   :  0.00   Min.   :    0.0  
##  1st Qu.: 2.38   1st Qu.:  0.95   1st Qu.:  0.05   1st Qu.:   41.0  
##  Median : 8.00   Median :  5.14   Median :  9.34   Median :   88.0  
##  Mean   :11.34   Mean   : 10.55   Mean   : 22.13   Mean   :  312.5  
##  3rd Qu.:19.88   3rd Qu.: 13.72   3rd Qu.: 34.91   3rd Qu.:  404.5  
##  Max.   :88.32   Max.   :100.00   Max.   :100.00   Max.   :38758.0  
##                                                    NA's   :83       
##  SaturatedFat.g.  Cholestrl_.mg.      Sugar.g.      Calcium.mg.    
##  Min.   : 0.000   Min.   :   0.00   Min.   : 0.000   Min.   :   0.00  
##  1st Qu.: 0.220   1st Qu.:   0.00   1st Qu.: 0.000   1st Qu.:  10.00  
##  Median : 1.592   Median :   4.00   Median : 1.840   Median :  21.00  
##  Mean   : 3.576   Mean   :  40.61   Mean   : 8.543   Mean   :  76.74  
##  3rd Qu.: 4.345   3rd Qu.:  67.00   3rd Qu.: 9.287   3rd Qu.:  69.00  
##  Max.   :95.600   Max.   :3100.00   Max.   :99.800   Max.   :7364.00  
##  NA's   :349      NA's   :410       NA's   :1832     NA's   :348      
##    Iron_.mg.      Potassium_.mg.     VitaminC.mg.       VitaminE.mg.    
##  Min.   :  0.00   Min.   :    0.0   Min.   :   0.000   Min.   :  0.000  
##  1st Qu.:  0.54   1st Qu.:  127.0   1st Qu.:   0.000   1st Qu.:  0.120  
##  Median :  1.38   Median :  229.5   Median :   0.000   Median :  0.300  
##  Mean   :  2.70   Mean   :  279.5   Mean   :   9.231   Mean   :  1.331  
##  3rd Qu.:  2.60   3rd Qu.:  336.0   3rd Qu.:   3.500   3rd Qu.:  0.800  
##  Max.   :123.60   Max.   :16500.0   Max.   :2732.000   Max.   :149.400  
##  NA's   :144      NA's   :426       NA's   :818        NA's   :2889     
##  VitaminD.microg.
##  Min.   :  0.000  
##  1st Qu.:  0.000  
##  Median :  0.000  
##  Mean   :  0.579  
##  3rd Qu.:  0.200  
##  Max.   :250.000  
##  NA's   :3262
```

```
## [1] "SALT,TABLE"
```

Thus, 100g of Salt has 38758mg of Sodium.

Let’s take a look at which foods have Sodium >3000mg (per 100g serving)

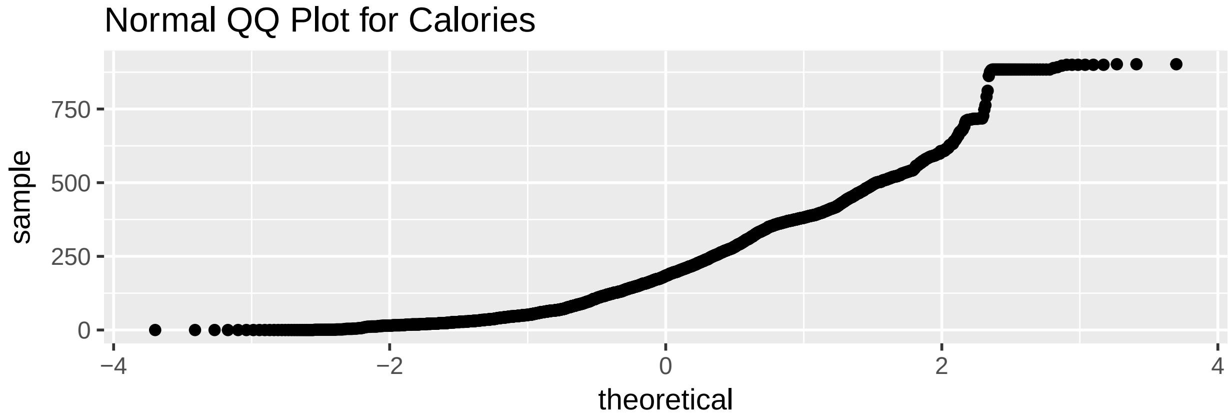
```
## [1] Dairy and Egg Products            Spices and Herbs                 
## [3] Soups, Sauces, and Gravies        Nut and Seed Products            
## [5] Finfish and Shellfish Products    Legumes and Legume Products      
## [7] Baked Products                    Sweets                           
## [9] Vegetables and Vegetable Products
## 25 Levels: American Indian/Alaska Native Foods ... Vegetables and Vegetable Products
```

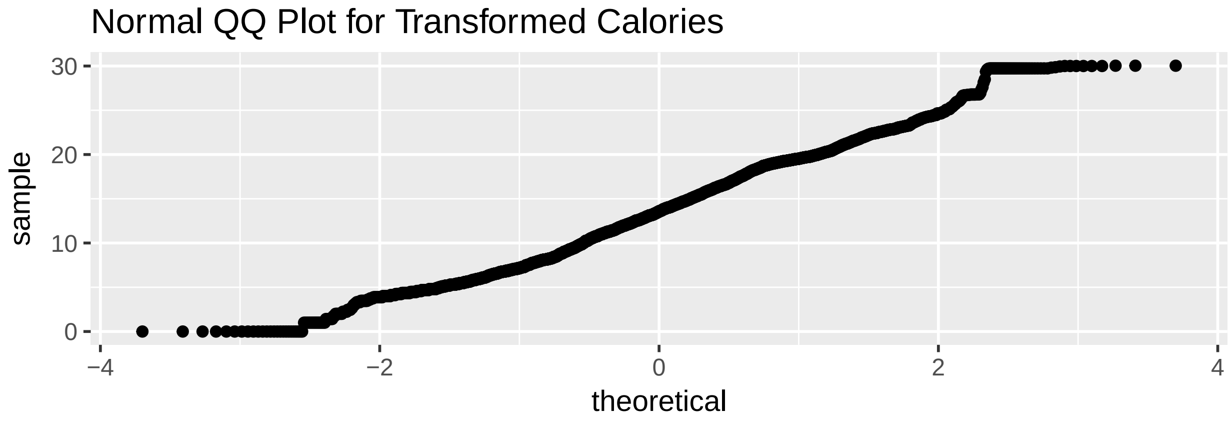
```
## 
##                    Baked Products            Dairy and Egg Products
##                                   3                                  2
##    Finfish and Shellfish Products       Legumes and Legume Products
##                                   4                                  5
##             Nut and Seed Products        Soups, Sauces, and Gravies
##                                   1                                 22
##                  Spices and Herbs                               Sweets
##                                   5                                  3
## Vegetables and Vegetable Products
##                                   2
```

Soups, Sauces, and Gravies is the group with the most items in High Sodium Category – (22)

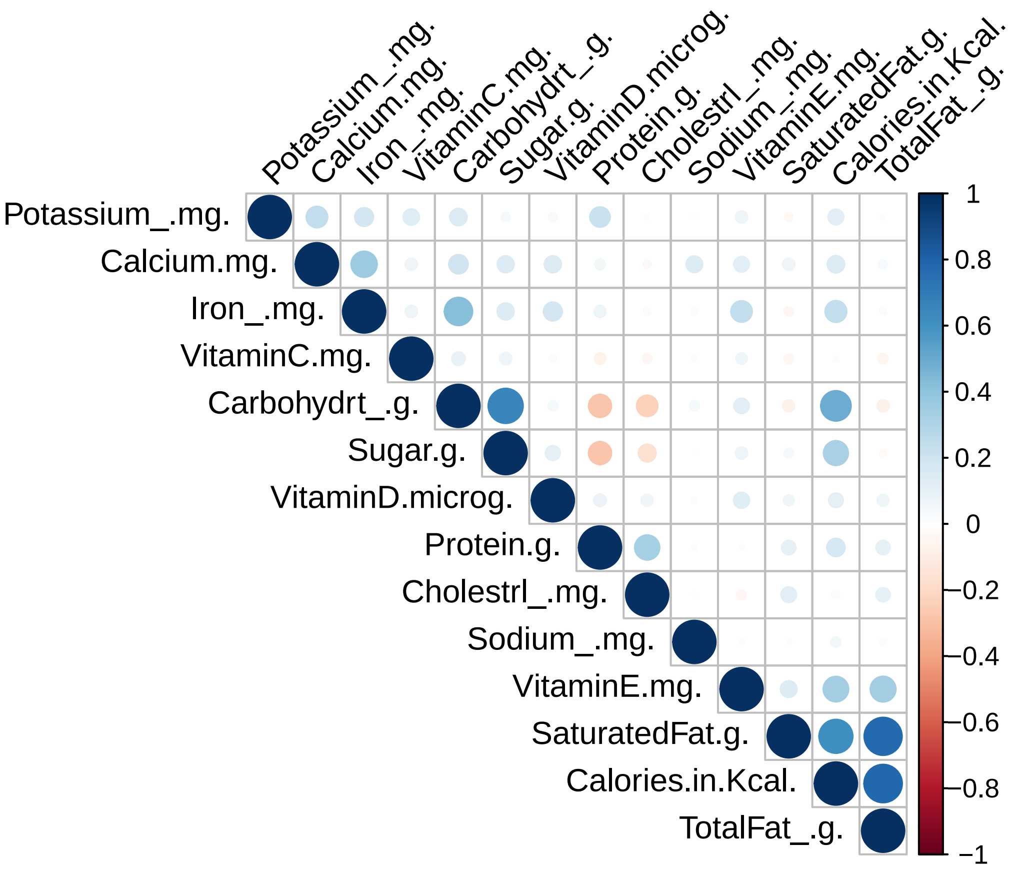
Let's Look at the normal probability plot for Calories

We take a look at how the distribution of Calories is spread across all food items. Since few food items contain higher calories than usual the density plot is skewed to the right.



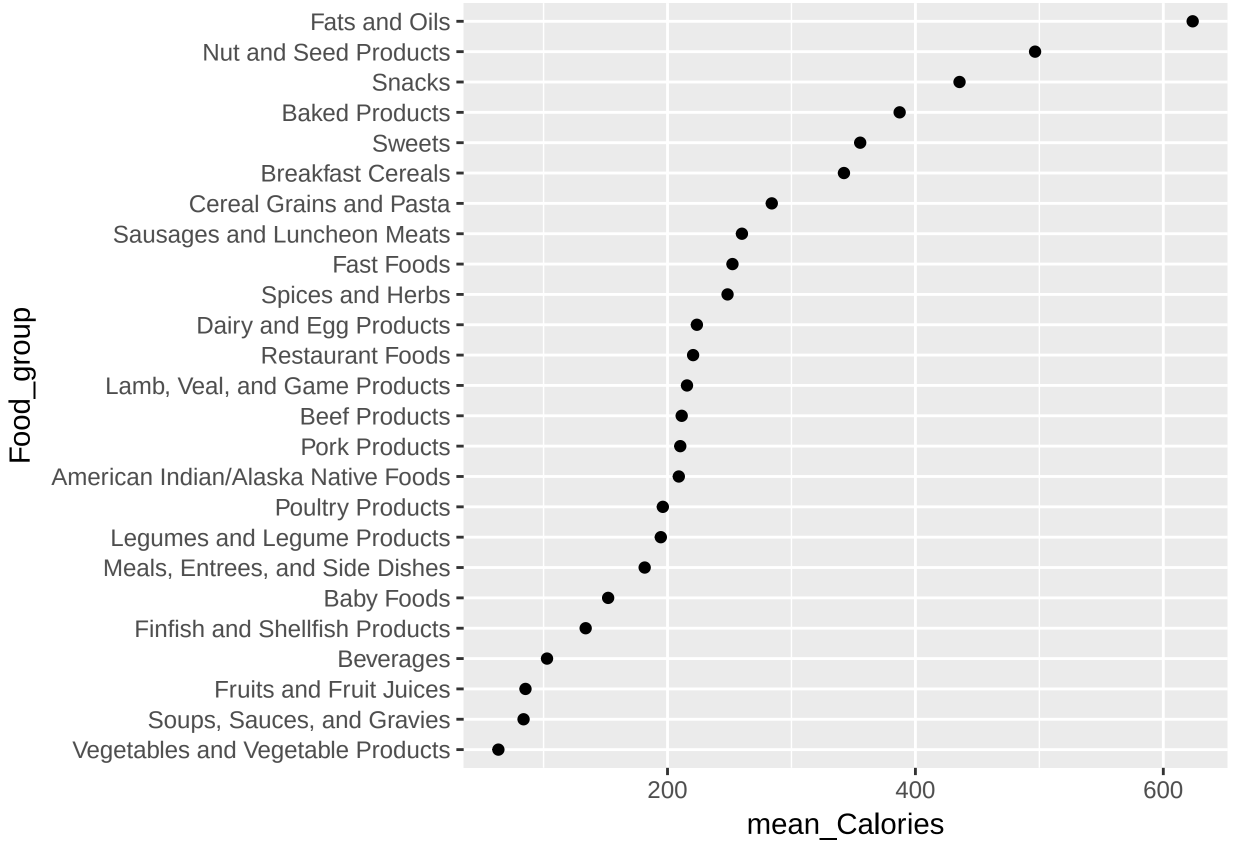


Correlation Among different variables



We can see that Calories is highly correlated with mainly Total Fats, Saturated Fats, Carbohydrates. Also, we can see that there is high correlation between Total Fats and saturated fats. Therefore, we can take any one of variable from Saturated and Total fats to built our linear model.Rest all our variables are not that strongly correlated.

The Mean calories with food group



From the plot we can see that most groups are located between 200 and 400 calories and these groups are mainly meals, meats, cereals and diary. Vegetables and fruits have much lower calories than other types of food and if people would like to lose weight and eat healthier, then vegetables and fruits are their best choices. On the other hand, keep away from fats and oils, snacks and baked products since the calories they contain are high. We define the level of high, medium and low calories are also highly depending on this graph.

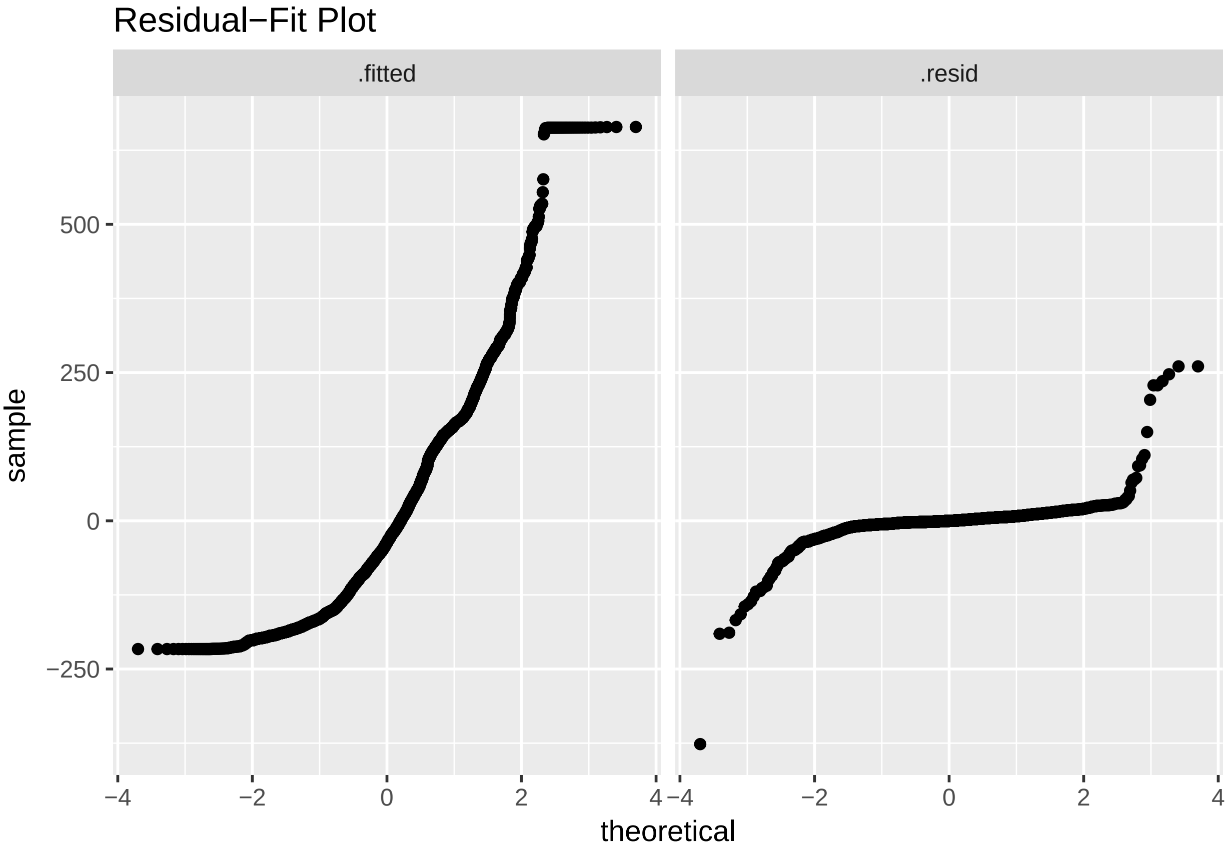
Regression Problem

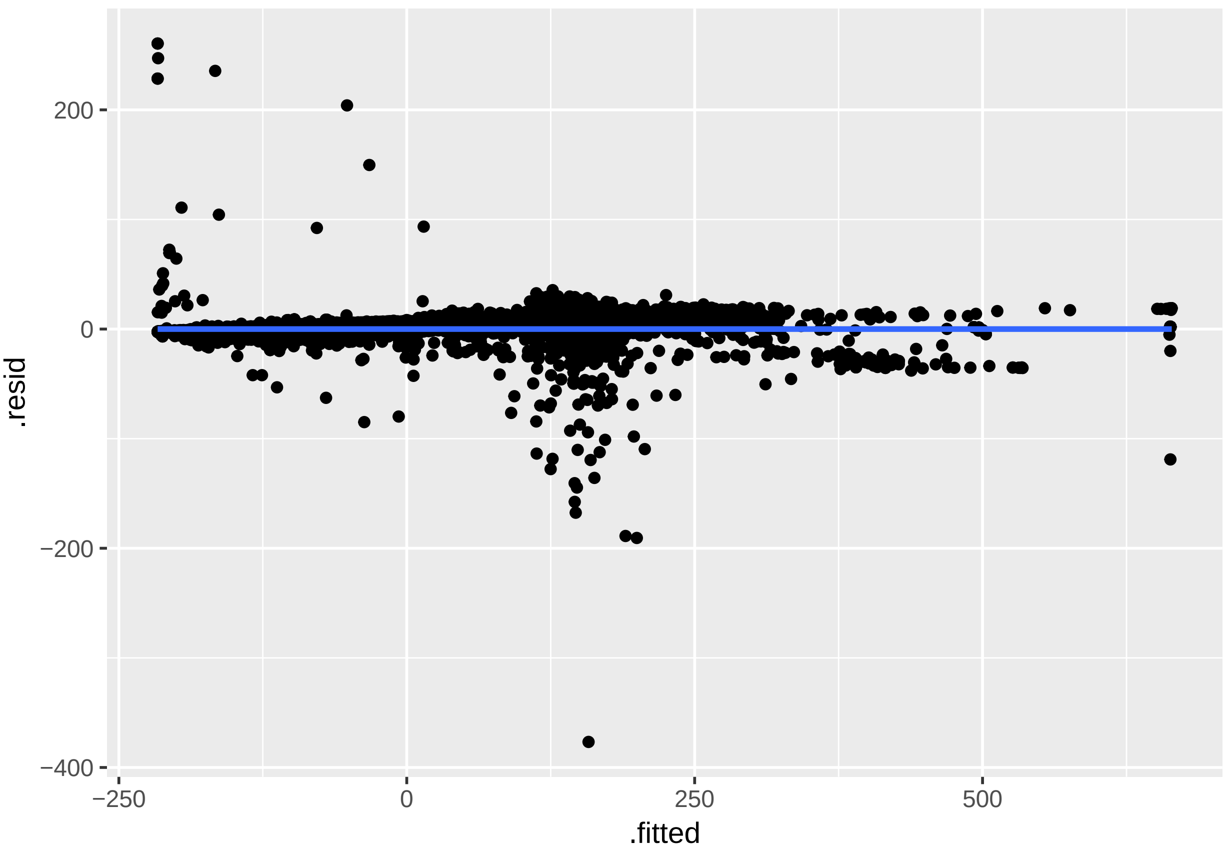
Linear Regression Model

Based on the correlation among variables, we build a regression model as follows- * **Dependent Variable** - Calories * **Predictor Variables** - Proteins, Total Fats, Carbohydrates, Cholestrol and Sugar

```
##
## Call:
## lm(formula = Calories.in.Kcal. ~ Protein.g. + TotalFat_.g. +
##     Carbohydrt_.g. + Sugar.g. + Cholestrl_.mg., data = model_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -376.68   -3.04   -0.16    5.37  260.50
##
## Coefficients:
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.500030   0.510888   4.893 1.02e-06 ***
## Protein.g.      4.015443   0.026411 152.038  < 2e-16 ***
## TotalFat_.g.    8.794368   0.017349 506.920  < 2e-16 ***
## Carbohydrt_.g. 3.741761   0.012900 290.066  < 2e-16 ***
## Sugar.g.        0.147862   0.023322   6.340 2.52e-10 ***
## Cholestrl_.mg. 0.010299   0.002782   3.702 0.000216 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.88 on 4616 degrees of freedom
## Multiple R-squared:  0.9886, Adjusted R-squared:  0.9885
## F-statistic: 7.972e+04 on 5 and 4616 DF,  p-value: < 2.2e-16
```





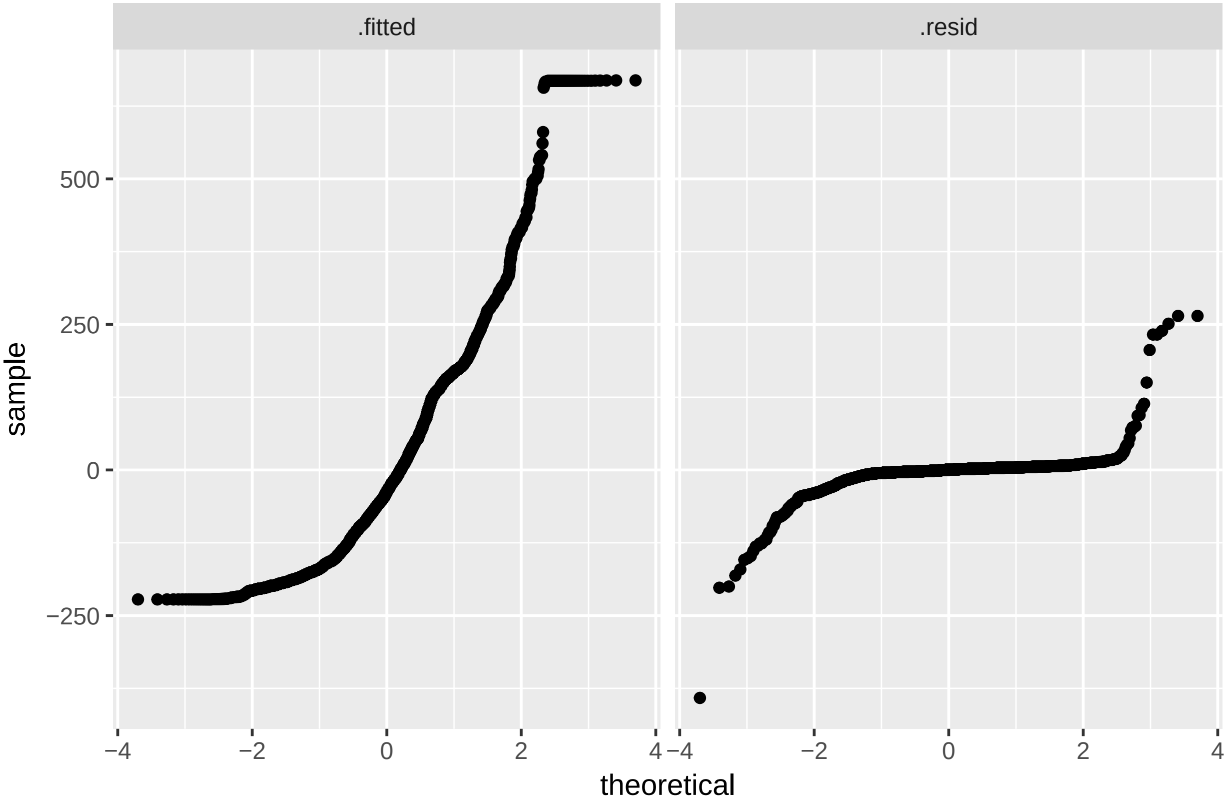
Linear model gives an R-squared of 98.85, which is very good. Looking at the residual fit plots we can see that there isn't much variation left in the residuals. The residuals are randonly distributed around 0 except for a few outliers.

All the predictor variables contribute to the total calories in a food item, some more than others (e.g. Total Fats).

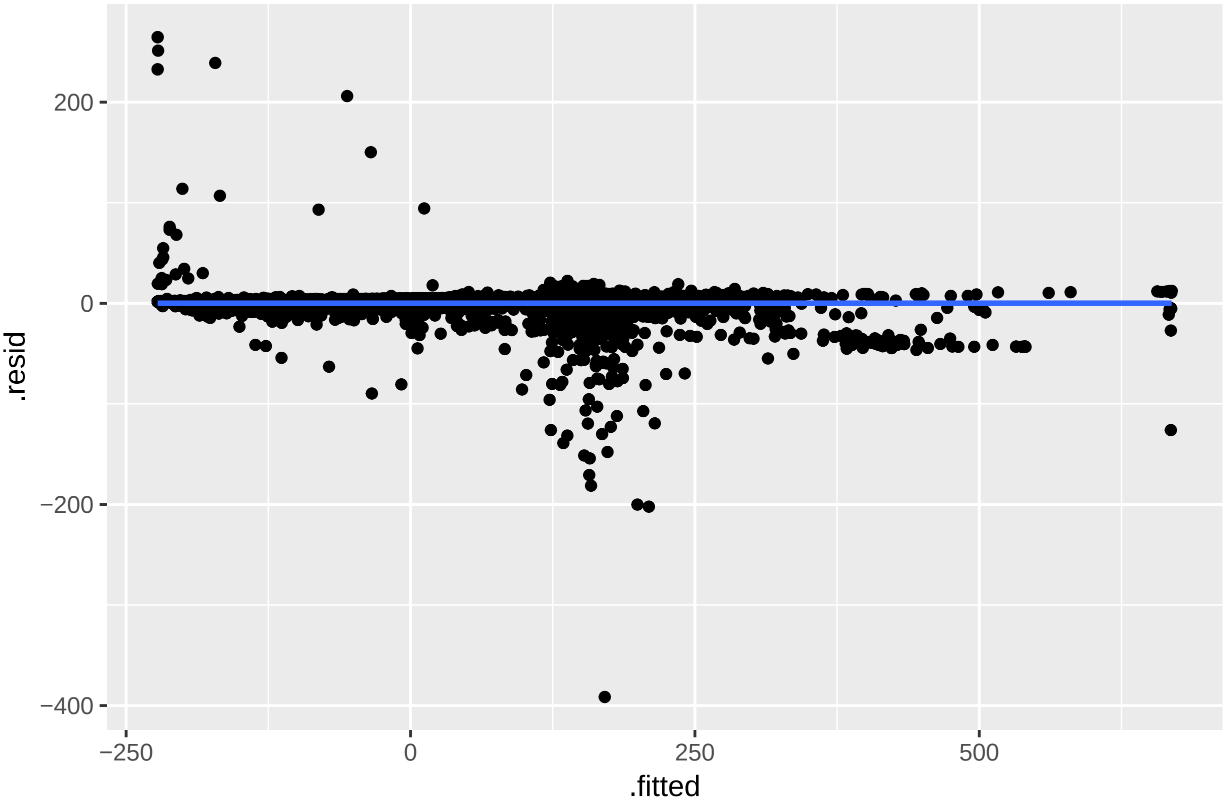
Robust linear model

We fit the robust linear model to see if the effect of outliers can be controlled.

Residual−Fit Plot



Residuals vs Fitted Plot



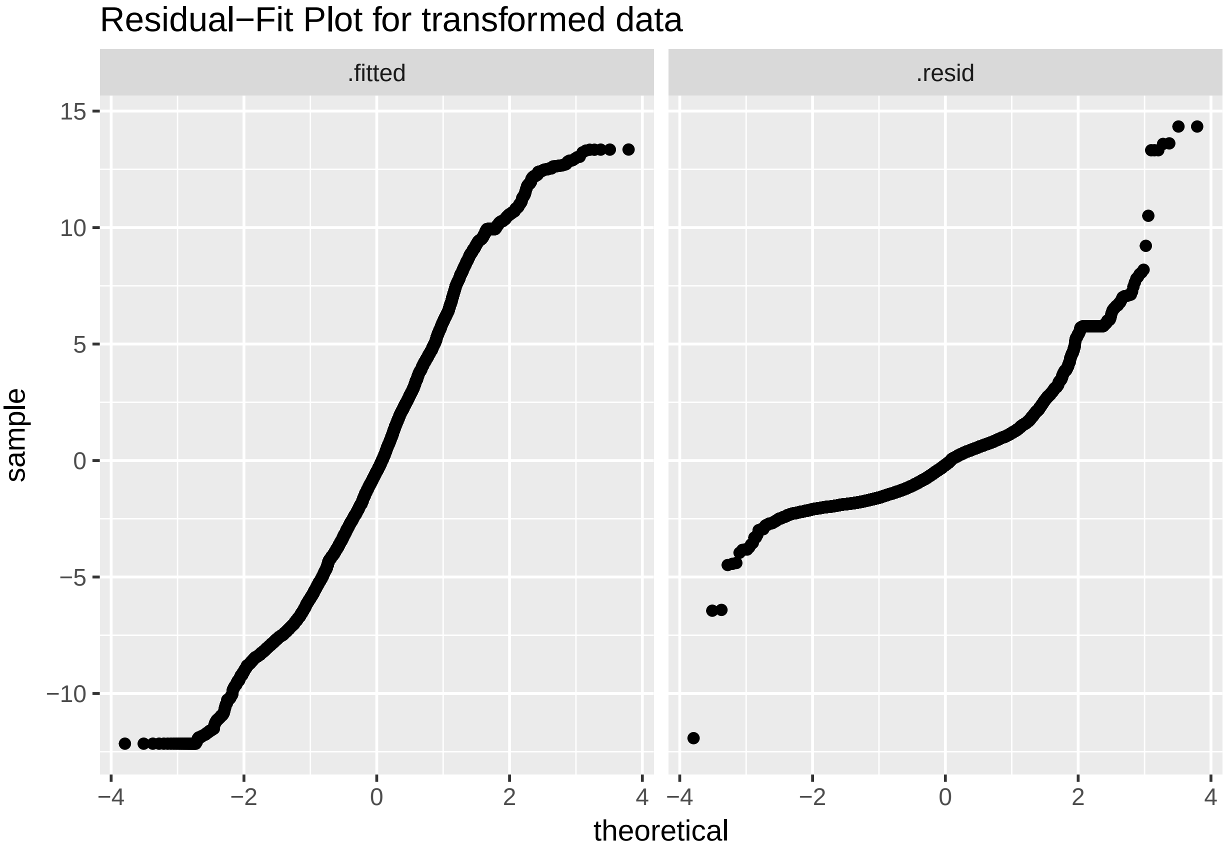
The plots of the robust linear model don't differ much from the linear model above and hence aren't included in our model.

Outliers

From the plots either in the linear model part or the Rubost part, several outliers exist and they may effect our model negatively. The outliers appear because for some particular food which contain high Calories but may contain low level of Protein, Fat, Carbohydrt, Sugar or Cholestrl or some low Calories foods may contain high level of Protein, Fat, Carbohydrt, Sugar or Cholestrl, these foods cause outliers because they are conflict with almost foods.

Transformations

Since the data for the variables under consideration is skewed right, we choose to transform the variables using **sqrt** transformation.



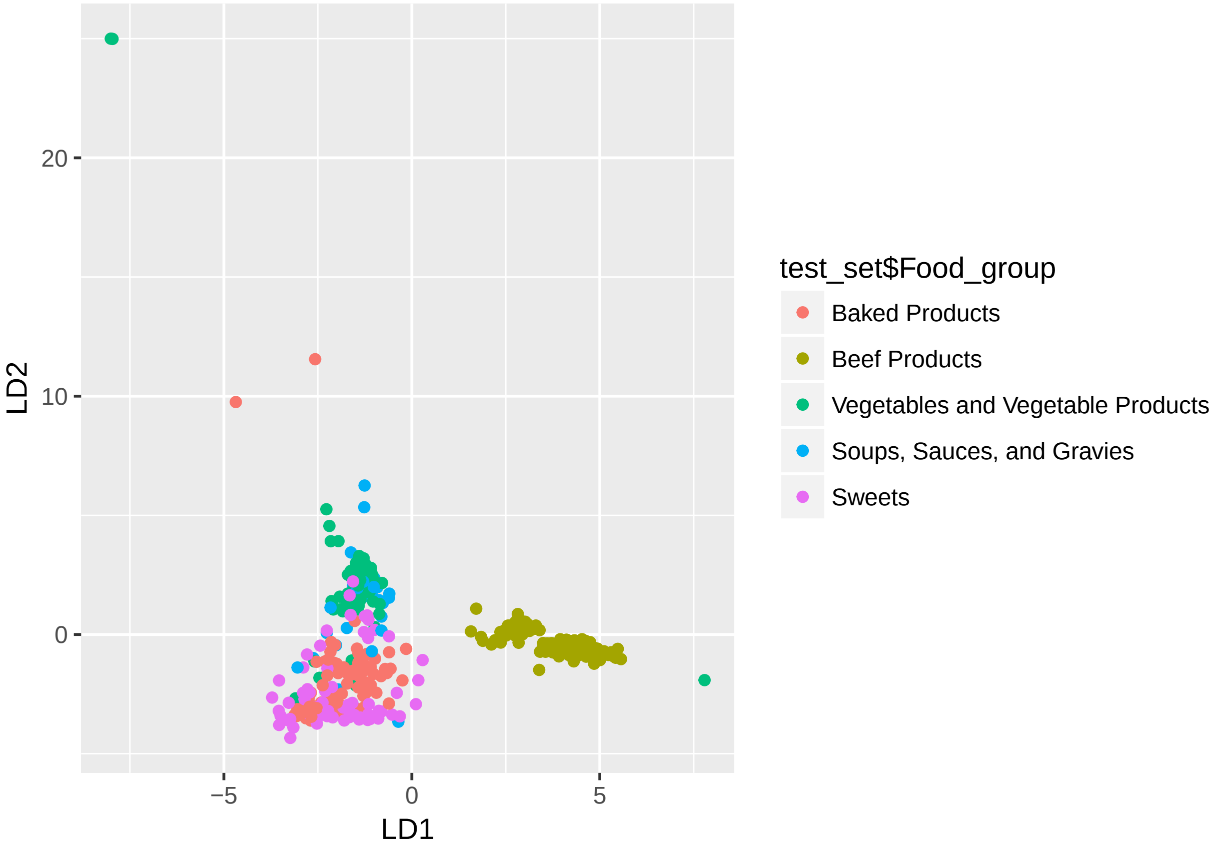
Classification Problem

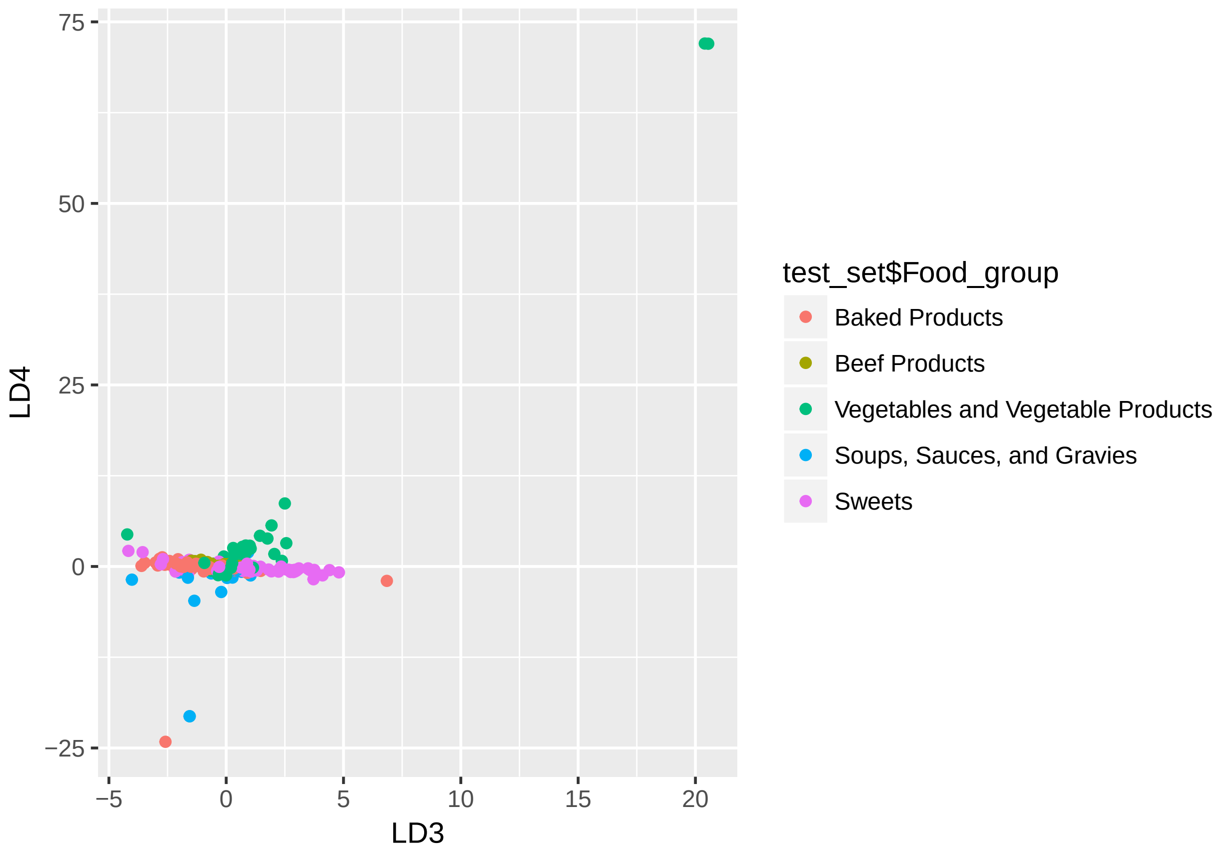
Question : Given a food item with nutrient content, can we predict the food categories to which it belong.

For classification problem, we have tries two different techniques: 1. LDA 2. Classification trees

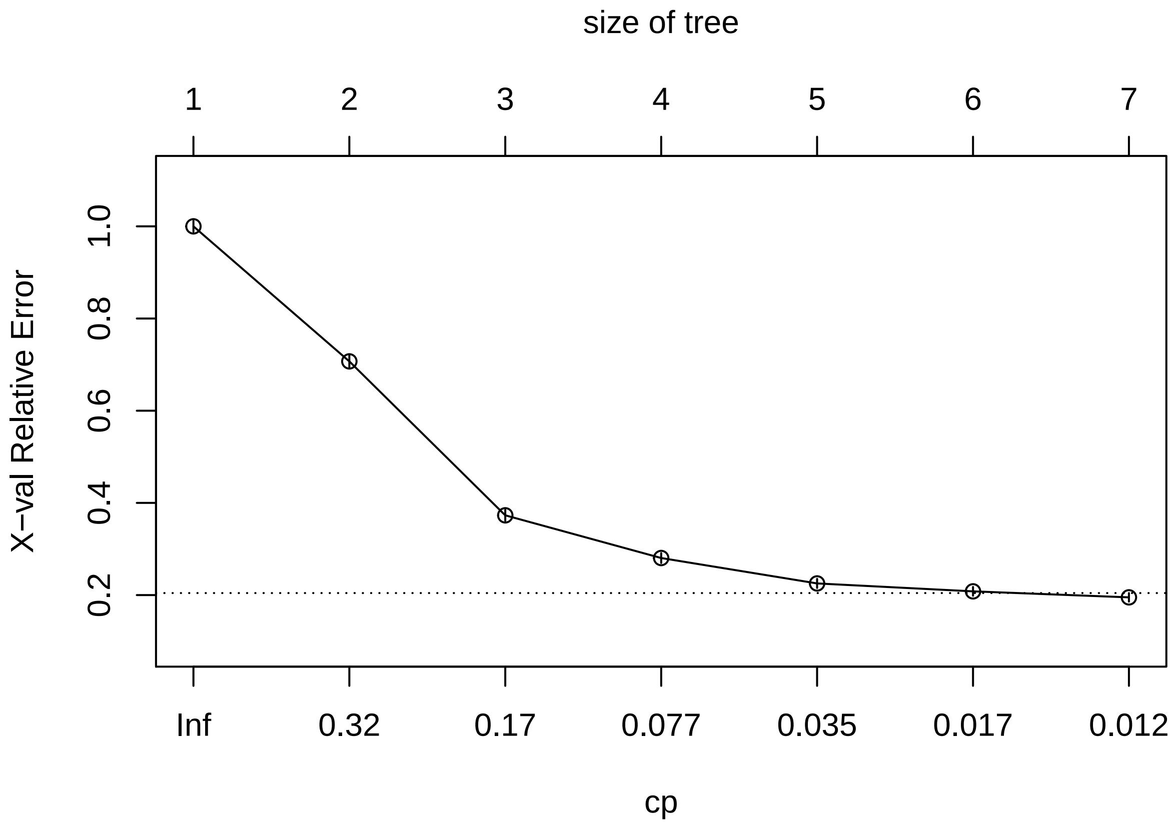
We extracted dataset of only five categories (Baked, Beef Products, Vegetables, Sopus and sweets) out of 25 categories and then divided our dataset into training set and testing set.

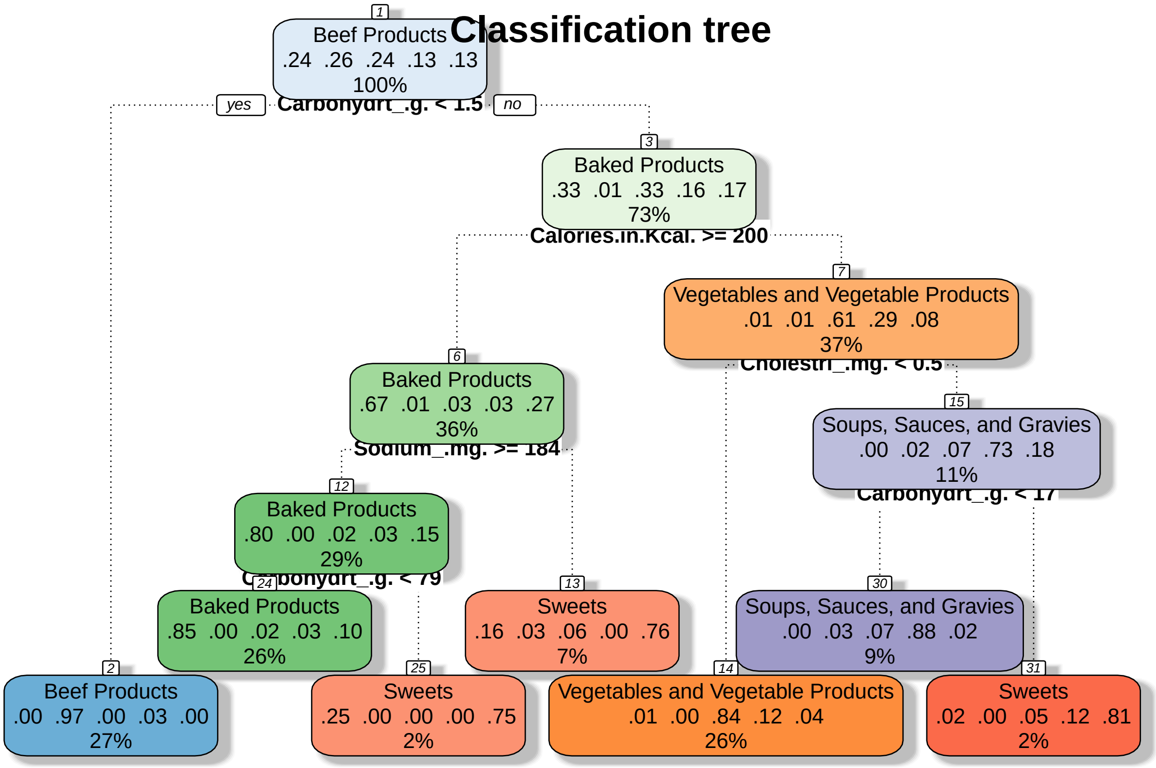
Our training set consists of 2610 examples and testing set contains 1000 examples. Since our main goal is classification, we tried to predict the category using LDA. As we know that the purpose of LDA is to find linear combinations of variables that gives the best possible spearation between the groups in our data.





As we can see that it is difficult to separate the groups using LDA as it is difficult to find linear separable boundaries in graph. Therefore we have tried classification trees to see if we can get good accuracy using that.


size of tree
1
2
3
4
5
6
7
X−val Relative Error
1.0
0.8
0.6
0.4
0.2
Inf
0.32
0.17
0.077
0.035
0.017
0.012
cp



Rattle 2017−Dec−13 19:45:19 sneha

Using classification trees we get approximately 89% accuracy on training set and 85% accuracy on tetsing set.Also, we see that carbohydrates are the main nutrient to classify the food item into different categories.

Future Work

We can explore some more interesting problems with the dataset.

1. Most of the food items we find in market has nutrient fact and ingredient list associated with it.Just with the nutrient data and ingredient list, can we find out the percentage of individual ingredient in the food item. We can also comment about the correctness of nutrient content.
2. Normally the food article only specify the general category of ingredient. For example, bread found in market does not specify the type of wheat they are using. We can comment about the type of wheat they ae using. This can help people with allergies about specific food particle.
3. By knowing the specific type of product, we can also comment whether a particular industry is importing a given food ingredient or not. If we have data on current trends of the cost of the food particular we are using then we may be able to predict the stocks prices of that particular industry.