Burger King

Miguel Dias PG40968

25/12/2019

Introduction

The following statistical analysis uses the dataset relating to the nutritional data of various hamburgers, appetizers, sides, and desserts served at the Burger King chain of restaurants.

It was intended to see the correlation that the amount of macronutrients such as protein, fat and sugar have in the total amount of calories in each food. In addition, it was intended to see whether the fact that the food had meat or was part of breakfast menus had an influence on the amount of calories.

```
setwd("C:/Users/Dias/Desktop")
burger = read.csv("burger_king.csv", sep = ";", dec = ".")
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.0.5
```

```
## corrplot 0.88 loaded
```

Exploratory data analysis:

Here is a summary of the values found for the minimum, maximum, mean, median as well as the values of the quadrants and the amount of data not available (NA).

The categorical variables were the presence of meat in the food or whether the food was part of a breakfast menu ("Meat" and "Breakfast" columns in the original dataset respectively) counting for the nominal groups "Yes" and "No". To represent the categorical date, it was important to use the as.factor function to define the "Yes" and "No" groups as integers and then perform the data regression.

```
summary(burger)
```

```
##    ï..Item           Serving.size      Calories         Fat.Cal
##  Length:122         Min.   : 43.0   Min.   :  25.0   Min.   :  0.0
##  Class :character   1st Qu.:113.5   1st Qu.: 310.0   1st Qu.:127.5
##  Mode  :character   Median :158.0   Median : 410.0   Median :190.0
##                     Mean   :167.7   Mean   : 452.1   Mean   :205.7
##                     3rd Qu.:216.5   3rd Qu.: 550.0   3rd Qu.:267.0
##                     Max.   :487.0   Max.   :1310.0   Max.   :650.0
##                     NA's   :11                       NA's   :6
##     Protein           Fat           Sat.Fat         Trans.fat
##  Min.   : 0.00   Min.   : 0.00   Min.   : 0.000   Min.   :0.0000
##  1st Qu.: 7.00   1st Qu.:14.25   1st Qu.: 4.000   1st Qu.:0.0000
##  Median :15.50   Median :22.00   Median : 7.000   Median :0.0000
```

```
##  Mean   :17.93   Mean   :24.78   Mean   : 8.701   Mean   :0.2992
##  3rd Qu.:24.75   3rd Qu.:33.00   3rd Qu.:12.000   3rd Qu.:0.5000
##  Max.   :71.00   Max.   :82.00   Max.   :32.000   Max.   :2.0000
##
##     Chol.mg.        Sodium.mg.         Carbs            Fiber
##  Min.   :  0.00   Min.   :   0.0   Min.   :  2.00   Min.   :0.000
##  1st Qu.: 20.00   1st Qu.: 490.0   1st Qu.: 27.00   1st Qu.:1.000
##  Median : 52.50   Median : 905.0   Median : 34.50   Median :1.000
##  Mean   : 92.75   Mean   : 918.8   Mean   : 39.37   Mean   :1.905
##  3rd Qu.:143.75   3rd Qu.:1245.0   3rd Qu.: 50.00   3rd Qu.:3.000
##  Max.   :455.00   Max.   :2490.0   Max.   :134.00   Max.   :9.000
##                                                     NA's   :6
##      Sugar            Meat            Breakfast          CarbsxMeat
##  Min.   : 0.000   Length:122         Length:122         Min.   :  0.00
##  1st Qu.: 3.000   Class :character   Class :character   1st Qu.:  0.00
##  Median : 6.000   Mode  :character   Mode  :character   Median : 26.50
##  Mean   : 9.975                                        Mean   : 23.48
##  3rd Qu.:10.000                                        3rd Qu.: 35.00
##  Max.   :58.000                                        Max.   :134.00
##
```

```
as.factor(burger$Meat)
```

```
##   [1] Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes
##  [19] Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes No
##  [37] Yes Yes Yes Yes Yes Yes No  No  No  No  No  No  No  No  No  No  No  No
##  [55] No  No  No  No  Yes Yes No  No  No  No  No  Yes Yes Yes No  No  Yes Yes
##  [73] Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes No  Yes No  No  No
##  [91] No  No  No  No  No  No  Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes No
## [109] No  No  No  No  No  No  No  No  No  No  No  No  No  No
## Levels: No Yes
```

```
as.factor(burger$Breakfast)
```

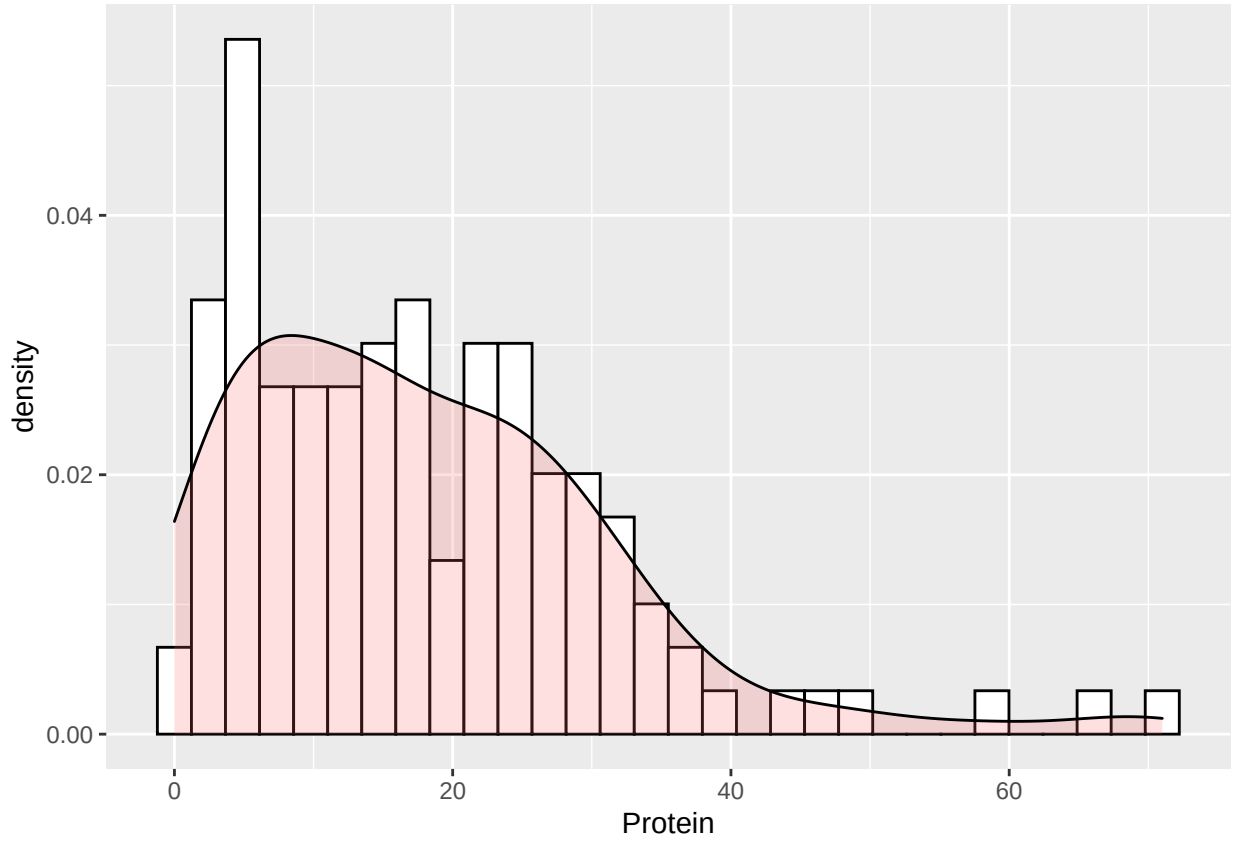
```
##   [1] No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No
##  [19] No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No
##  [37] No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No  No
##  [55] No  No  No  No  No  No  No  No  Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes
##  [73] Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes
##  [91] Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes No
## [109] No  No  No  No  No  No  No  No  No  No  No  No  No  No
## Levels: No Yes
```

Quantitative variables

Histograms were made with the main variables to be used, with the continuous dependent variable corresponding to the column of calories and the independent variables the amount of protein, fat and carbohydrates. In this way, it is possible to visualize the number of foods that fall within a macronutrient quantification interval.

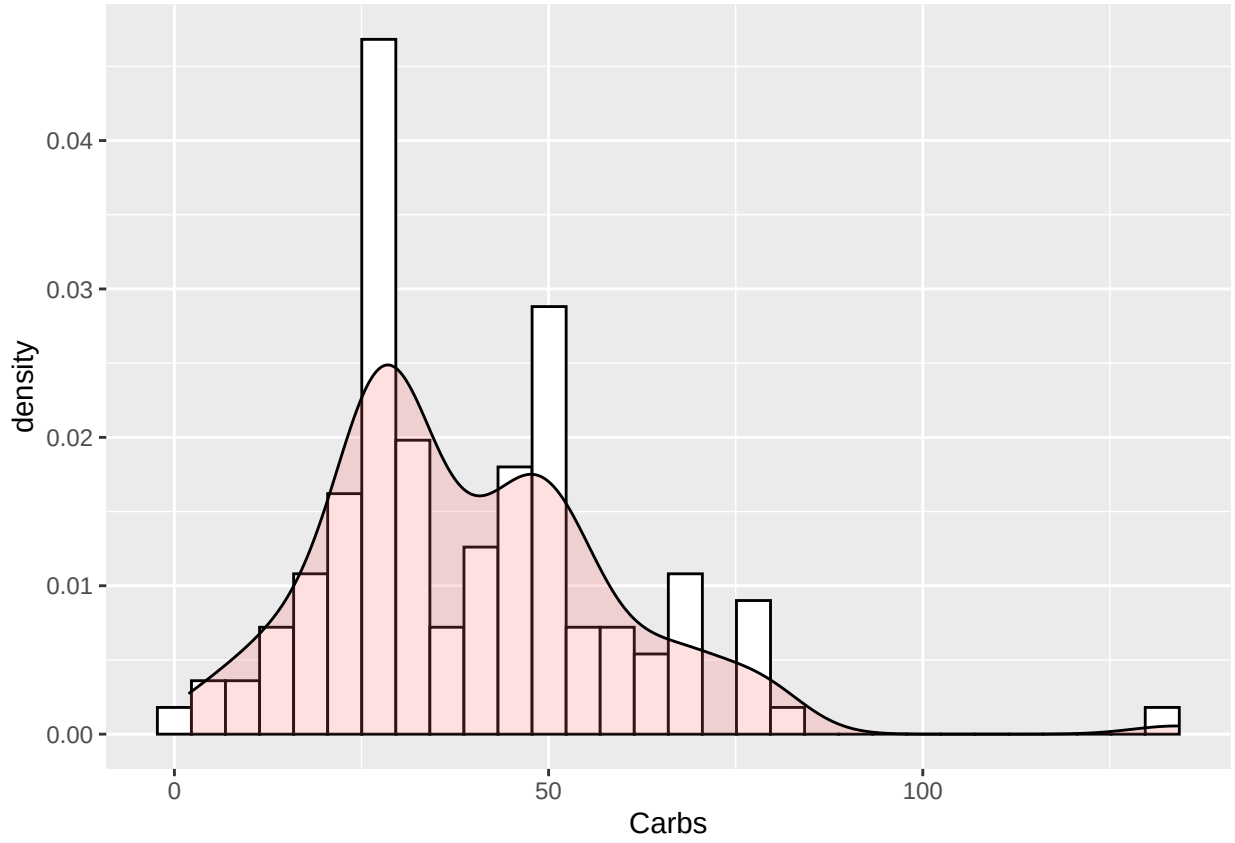
```
#Protein
ggplot(burger, aes(x=Protein)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")
```

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



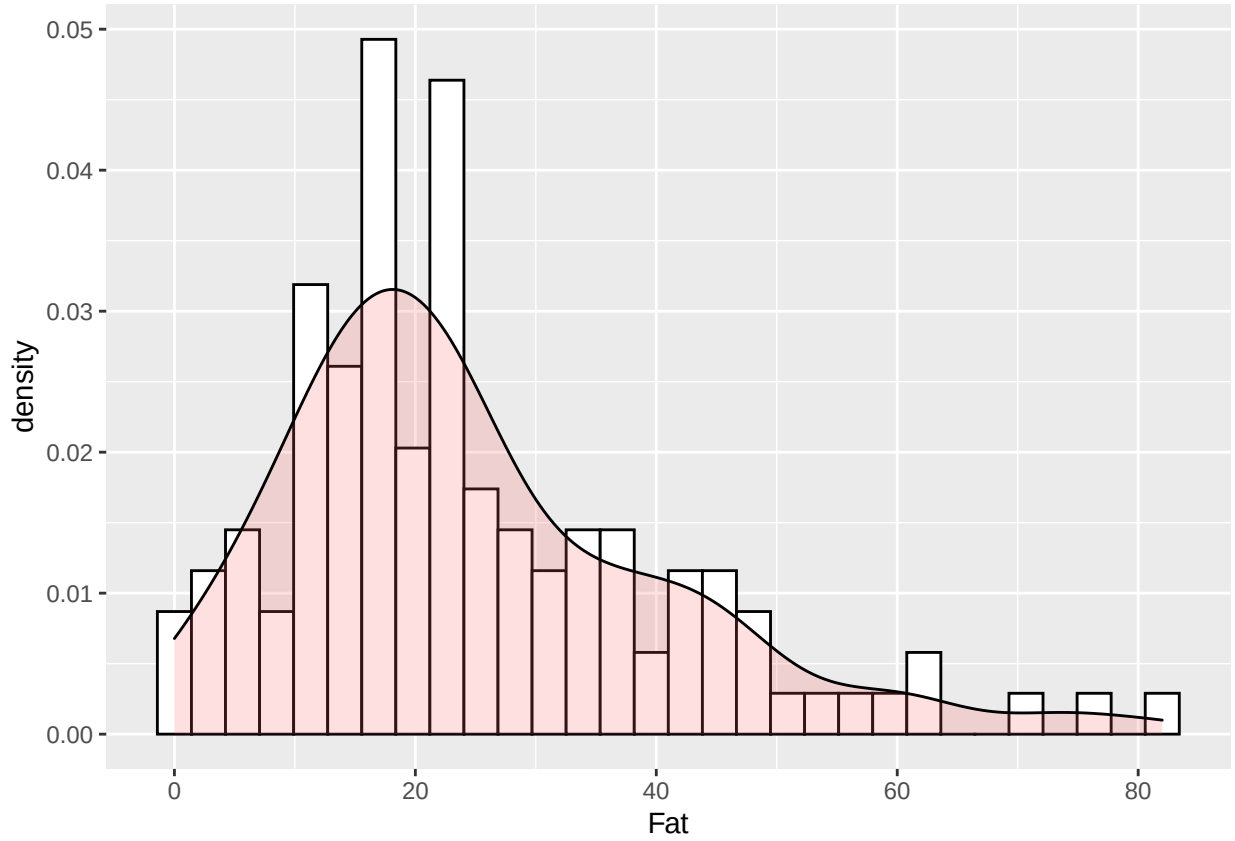

```
#Carbs
ggplot(burger, aes(x=Carbs)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")
```

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



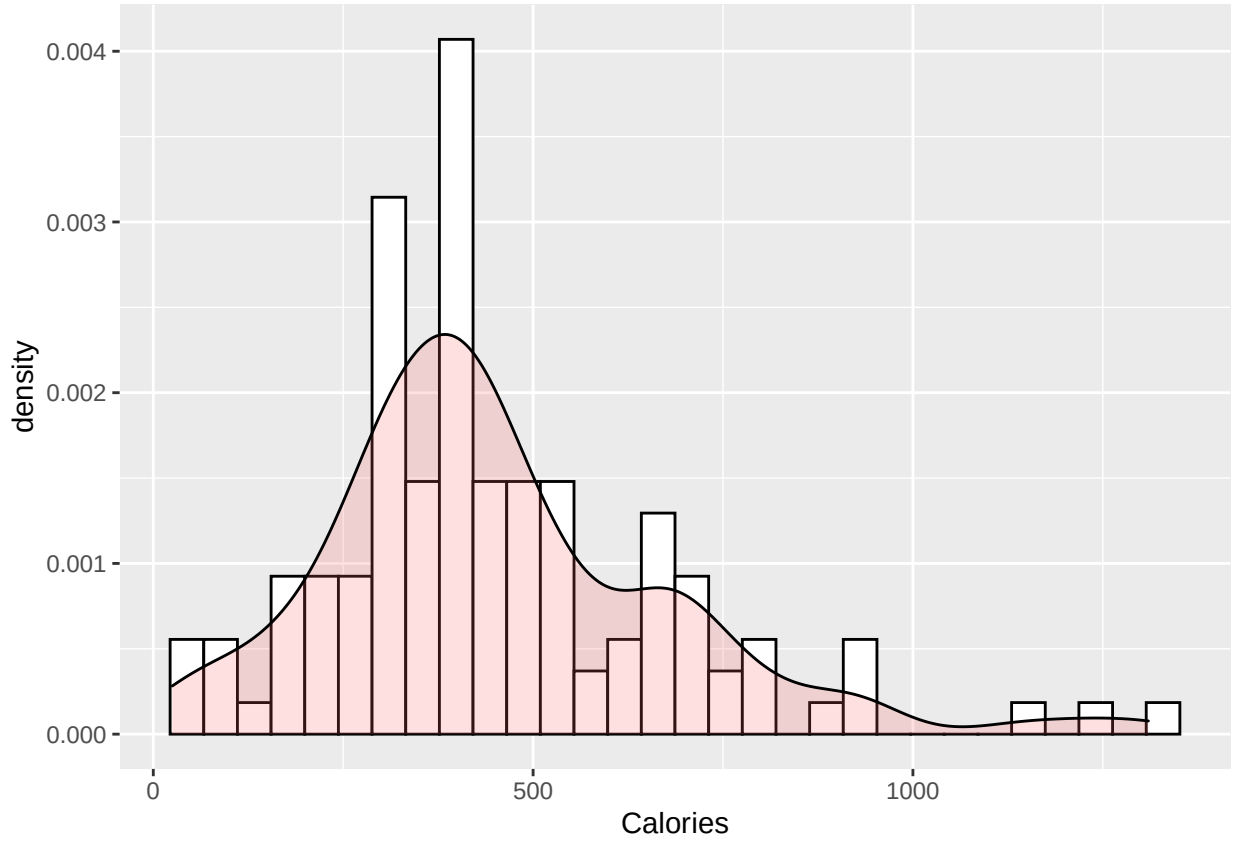

```
#Fat
ggplot(burger, aes(x=Fat)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")
```

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




```
#Calories
ggplot(burger, aes(x=Calories)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")
```

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



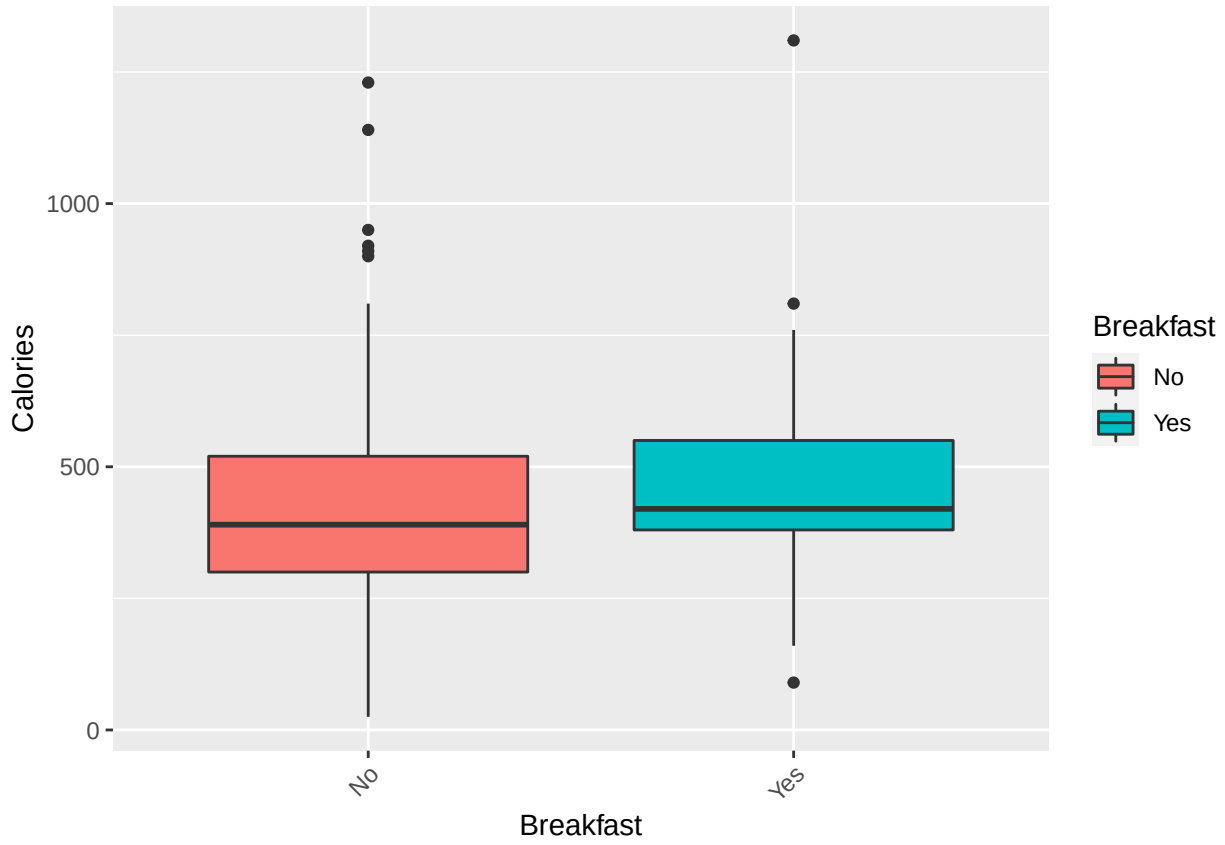
It is possible to see that most of the foods mostly have an amount between 250 and 500 calories, 5 to 40 grams of protein, 25 to 50 grams of carbohydrates and between 10 and 30 grams of fat respectively for each histogram represented. There are 3 foods that have a relatively high amount of proteins, hydrates and fat as well as a high caloric content being represented by the bars located further to the right of the density function graphs.

The functions do not follow a normal distribution, and for the hydrate function there are two maximum peaks. This may indicate the existence of mixing distributions of other types.

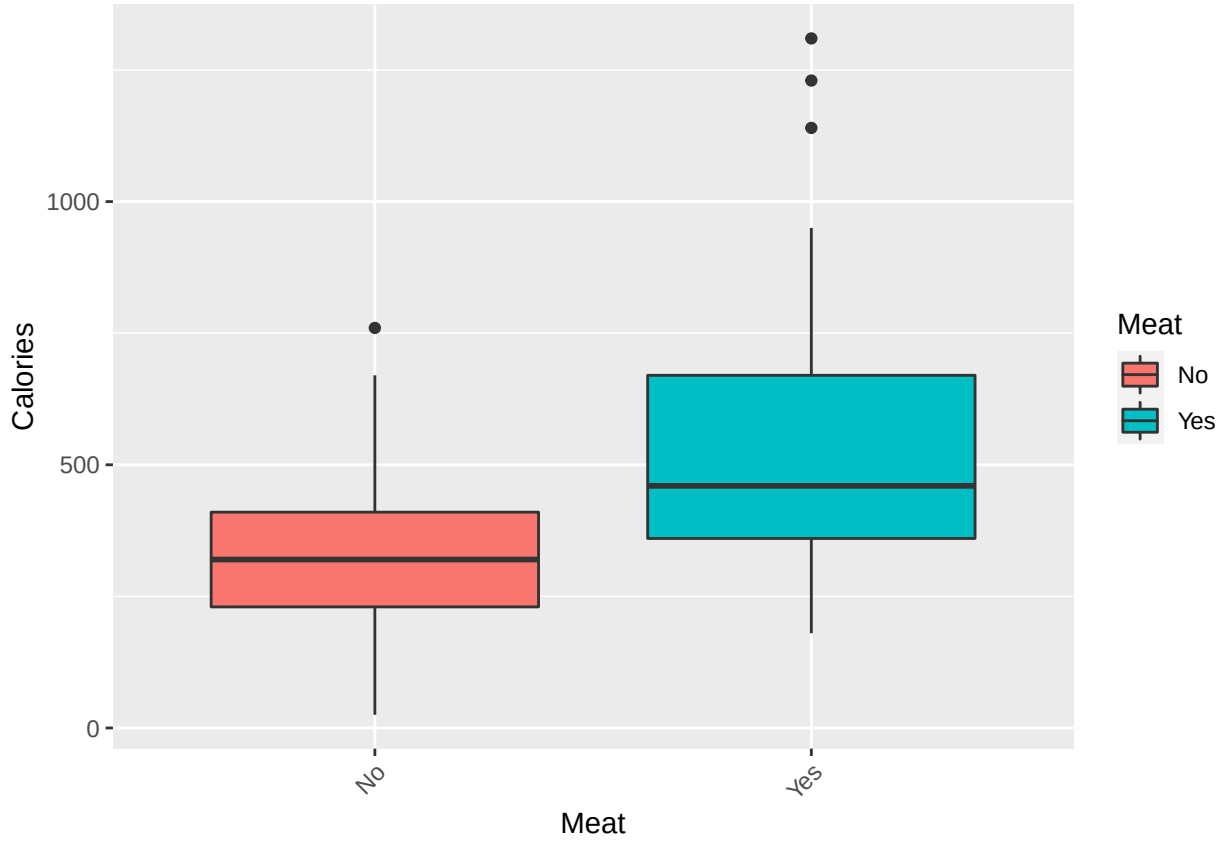
Categorical variables

For this type of variables, boxplots were made to characterize each nominal category. It was seen how the amount of calories varies depending on the type of food (if the food is breakfast) or if the food had meat.

```
ggplot(burger, aes(Breakfast, Calories)) +
  geom_boxplot(aes(fill = Breakfast)) +
  theme(axis.text.x = element_text(angle = 45, hjust =1)) + scale_x_discrete(name = 'Breakfast')
```




```
ggplot(burger, aes(Meat, Calories)) +
  geom_boxplot(aes(fill = Meat)) +
  theme(axis.text.x = element_text(angle = 45, hjust =1)) + scale_x_discrete(name = 'Meat')
```



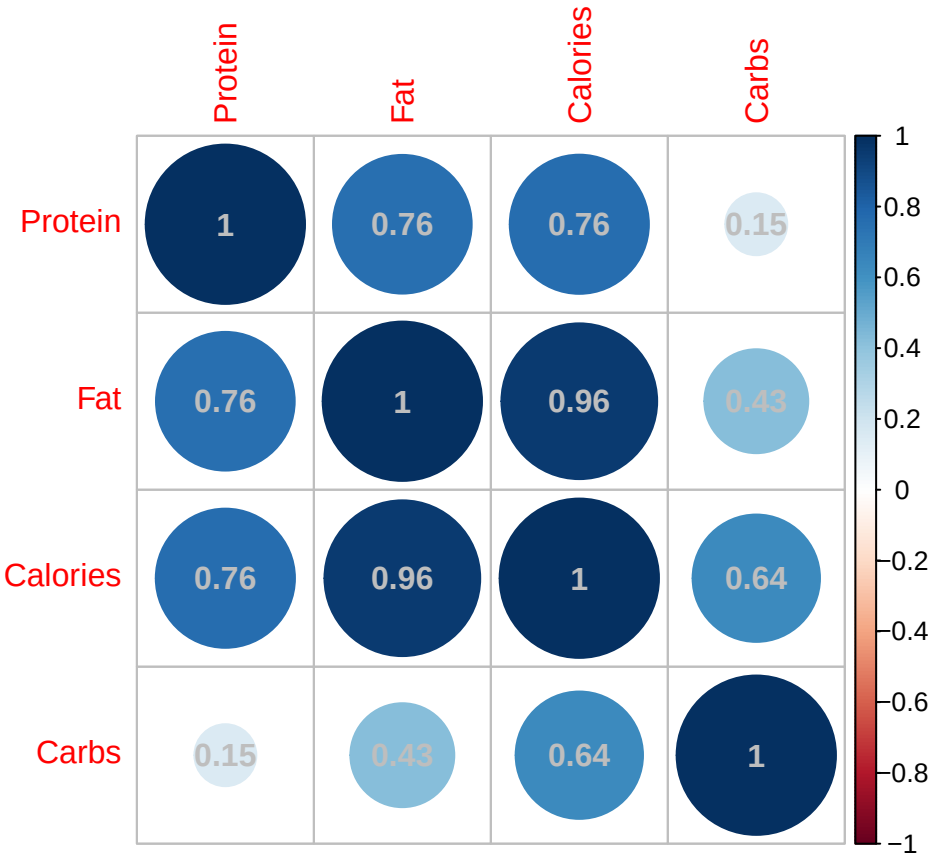
For the breakfast variable, we see that the median is approximately the same for both categories. It is also possible to see that the interquartile range is slightly higher for foods that are not considered breakfast and in the case of foods that are breakfast there is a value for the third quartile slightly higher, however the foods classified as not being breakfast show more outliers and the maximum caloric value obtained is slightly higher in comparison.

In the case of the boxplot created for the presence of meat, the most salient result is the greater width of the boxplot obtained for the "Yes" category and the fact that the median, interquartile range, maximum value and outliers are all higher in terms of calories compared to non-meat food. From a nutritional point of view this makes sense because meat is a very nutritionally dense food with a higher amount of fat, the most energetic macronutrient [reference: https://www.nal.usda.gov/fnic/how-many -calories-are-one-gram-fat-carbohydrate-or-protein].

Correlation of calories, protein, fat, carbohydrates:

A corrplot was performed to determine the degree of correlation between all the variables under study.

```
subc<-subset(burger,select = c(Calories,Protein ,Fat,Carbs))
M<-cor(subc)
corrplot(M, order = "AOE", addCoef.col = "grey")
```



It is possible to verify that the calorie variable presents the highest correlation with the fat variable (values close to one mean high correlation). Considering that 1 gram of fat has 9 calories[https://www.nal.usda.gov/fnic/how-many-calories-are-one-gram-fat-carbohydrate-or-protein], approximately twice as many calories present in the same amount of proteins or hydrates, this result is in line with expectations.

On the other hand, the pairs of variables hydrates and protein as well as fat and hydrates show low correlation. This could be explained by the fact that the molecules are very different, however, the protein and fat pair has a correlation value of 0.76 so this could not be the reason. Another reason that could lead to the high correlation value between fat and protein is the fact that meat is mainly made up of these macronutrients (and a low percentage of carbohydrates) and this catering chain mainly serves food with meat.

Linear model of the dependent variable

The results relative to the linear model as well as the sketched graphs are shown below.

```
yhat=lm(Calories ~ Protein + Carbs + Fat + Meat + Breakfast,data = burger)
summary(yhat)
```

```
##
## Call:
## lm(formula = Calories ~ Protein + Carbs + Fat + Meat + Breakfast,
##     data = burger)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.161  -3.993  -0.262   3.068  40.616
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.07981    2.04711  -0.039 0.968967
## Protein       4.03484    0.10716  37.651  < 2e-16 ***
## Carbs         3.88753    0.04541  85.618  < 2e-16 ***
## Fat           9.12411    0.08849 103.109  < 2e-16 ***
## MeatYes      -2.65253    1.99610  -1.329 0.186502
## BreakfastYes  6.13604    1.73424   3.538 0.000581 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.374 on 116 degrees of freedom
## Multiple R-squared:  0.9987, Adjusted R-squared:  0.9987
## F-statistic: 1.836e+04 on 5 and 116 DF,  p-value: < 2.2e-16
```

The variables corresponding to the macronutrients have very low p-value levels, indicating a high statistical significance of these parameters in the model. For categorical variables, breakfast has a p-value less than 0.05 so it is statically relevant to the model.

The p-value for meat, however, is greater than 0.05 and the nominal category "Yes" presents an estimation with a negative value. This negative value means that the values of the variable under study do not contribute to the dependent variable, which is in disagreement with what was observed in the boxplots outlined above. However, the high R-squared value of 0.9987 means that the placed parameters explain almost 100% of the variation in calories, so the inclusion of the meat variable could be included without compromising the integrity of the model. Other data that support the linear model are the elevated F-test statistic value and the corresponding reduced p-value.

Stepwise regression

In order to better validate the chosen linear model, a "stepwise" regression was used. This is a method of adjusting regression models in which the choice of variables for model prediction is performed by an automatic process using an algorithm or add the variables one by one, depending on the chosen direction, to find the most suitable model.

This type of regression uses an error estimator called the Akaike Information Criterion (AIC) that measures the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. The smaller the calculated AIC value, the better the model.

Backward

First, this regression was carried out with the "backward" direction, with the results shown below. In this direction the AIC value is calculated taking into account all the variables and then removes the variables that decrease the AIC value.

```
step_b<-step(yhat,direction = "backward")
```

```
## Start:  AIC=524.37
## Calories ~ Protein + Carbs + Fat + Meat + Breakfast
##
##             Df Sum of Sq    RSS     AIC
## - Meat       1       124   8258  524.21
## <none>                     8134  524.37
## - Breakfast  1       878   9011  534.87
## - Protein    1     99399 107533  837.35
## - Carbs      1    514002 522136 1030.12
```

```
## - Fat        1    745454 753588 1074.89
##
## Step:  AIC=524.21
## Calories ~ Protein + Carbs + Fat + Breakfast
##
##             Df Sum of Sq    RSS     AIC
## <none>                     8258  524.21
## - Breakfast  1       781   9039  533.24
## - Protein    1    116859 125116  853.82
## - Carbs      1    525753 534011 1030.87
## - Fat        1    745411 753668 1072.90
```

This regression started with an AIC value of 524.37, which was reduced to 524.21 by removing the meat variable. Of all the parameters offered, this was the most disposable and the only one that did not compromise the viability of the model in any way, which goes somewhat against the fact that in the previously performed regression, this parameter received a p-value greater than 0.05. However, it is important to note that the difference between including this variable in the study or not is extremely minimal and if a statistician wanted to include it in the linear model, the validity of the model would not suffer much.

The same that was said now can be said for the breakfast variable, the difference between including it in the model or not is in the hands of the person who deals with these results, in this case if it were removed the AIC value would rise to 533.24, an increase of only 7 points.

The variables that correspond to the macronutrients, in turn, are too important for this study, if they are removed the AIC value will increase significantly and the model will lose quality.

Forward

Regressions with this direction add variables in order to make the AIC value go down. The result for the AIC value in this case was equal to 524.37, maintaining the categorical variable for meat.

```
step_f<-step(yhat,direction = "forward")
```

```
## Start:  AIC=524.37
## Calories ~ Protein + Carbs + Fat + Meat + Breakfast
```

Both

Finally, a regression was performed considering both directions. At the beginning, the regression calculated the AIC with all the variables being the value 524.37 and in the second part it removed the meat variable and the AIC value went down again to 524.21. In the stepwise results it was shown that the addition of meat in the model would increase the AIC value back to the starting value.

```
step_bo<-step(yhat,direction = "both")
```

```
## Start:  AIC=524.37
## Calories ~ Protein + Carbs + Fat + Meat + Breakfast
##
##             Df Sum of Sq    RSS     AIC
## - Meat       1       124   8258  524.21
## <none>                     8134  524.37
## - Breakfast  1       878   9011  534.87
## - Protein    1     99399 107533  837.35
## - Carbs      1    514002 522136 1030.12
## - Fat        1    745454 753588 1074.89
##
## Step:  AIC=524.21
## Calories ~ Protein + Carbs + Fat + Breakfast
```

```
##
##             Df Sum of Sq    RSS     AIC
## <none>                     8258  524.21
## + Meat       1       124   8134  524.37
## - Breakfast  1       781   9039  533.24
## - Protein    1    116859 125116  853.82
## - Carbs      1    525753 534011 1030.87
## - Fat        1    745411 753668 1072.90
```

It is thus concluded that the inclusion of all variables follows an optimal linear model for data analysis.

Verification of waste conditions

Now you can check the conditions of the residuals using the linear model after going through the stepwise algorithm. To avoid redundancy, only the graphs of the conditions of the residuals related to the "stepwise" regression with the "backward" direction were presented. Below is the regression without using the algorithm (that is, it includes all variables) and with the stepwise algorithm.

```
summary(yhat)
```

```
##
## Call:
## lm(formula = Calories ~ Protein + Carbs + Fat + Meat + Breakfast,
##     data = burger)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.161  -3.993  -0.262   3.068  40.616
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.07981    2.04711  -0.039 0.968967
## Protein       4.03484    0.10716  37.651  < 2e-16 ***
## Carbs         3.88753    0.04541  85.618  < 2e-16 ***
## Fat           9.12411    0.08849 103.109  < 2e-16 ***
## MeatYes      -2.65253    1.99610  -1.329 0.186502
## BreakfastYes  6.13604    1.73424   3.538 0.000581 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.374 on 116 degrees of freedom
## Multiple R-squared:  0.9987, Adjusted R-squared:  0.9987
## F-statistic: 1.836e+04 on 5 and 116 DF,  p-value: < 2.2e-16
```

```
summary(step_b)
```

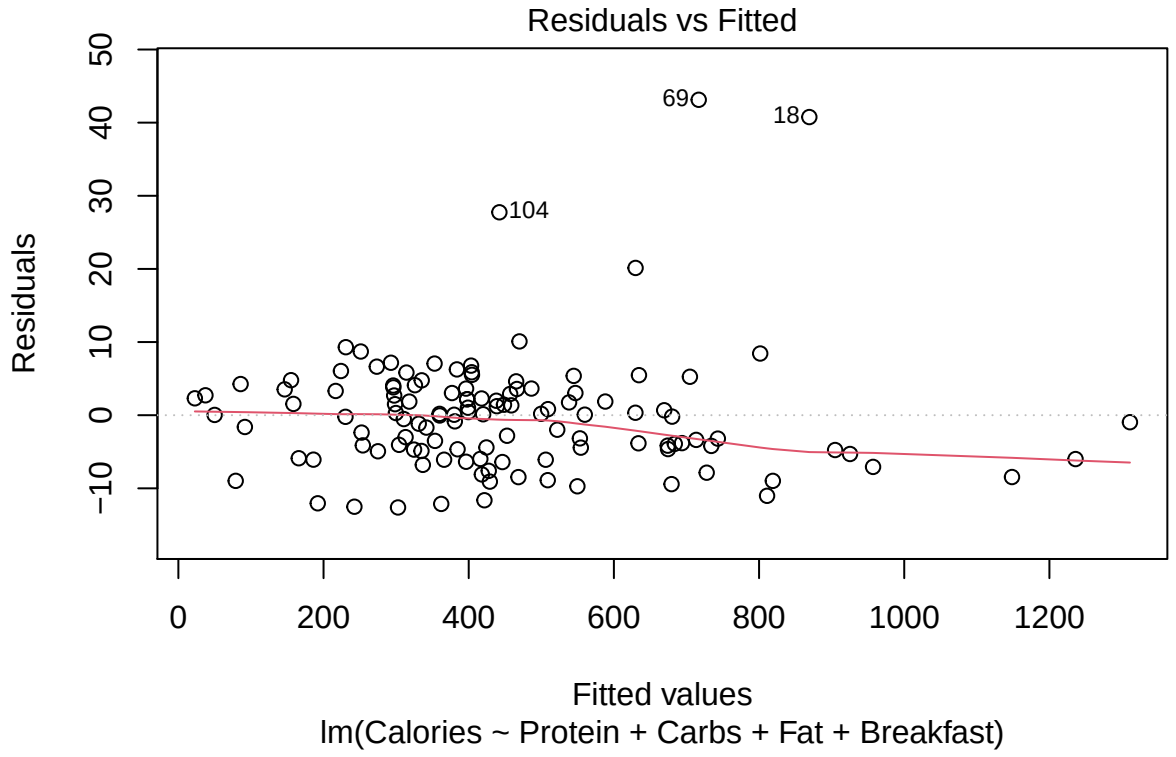
```
##
## Call:
## lm(formula = Calories ~ Protein + Carbs + Fat + Breakfast, data = burger)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.614  -4.747   0.050   3.478  43.124
##
## Coefficients:
```

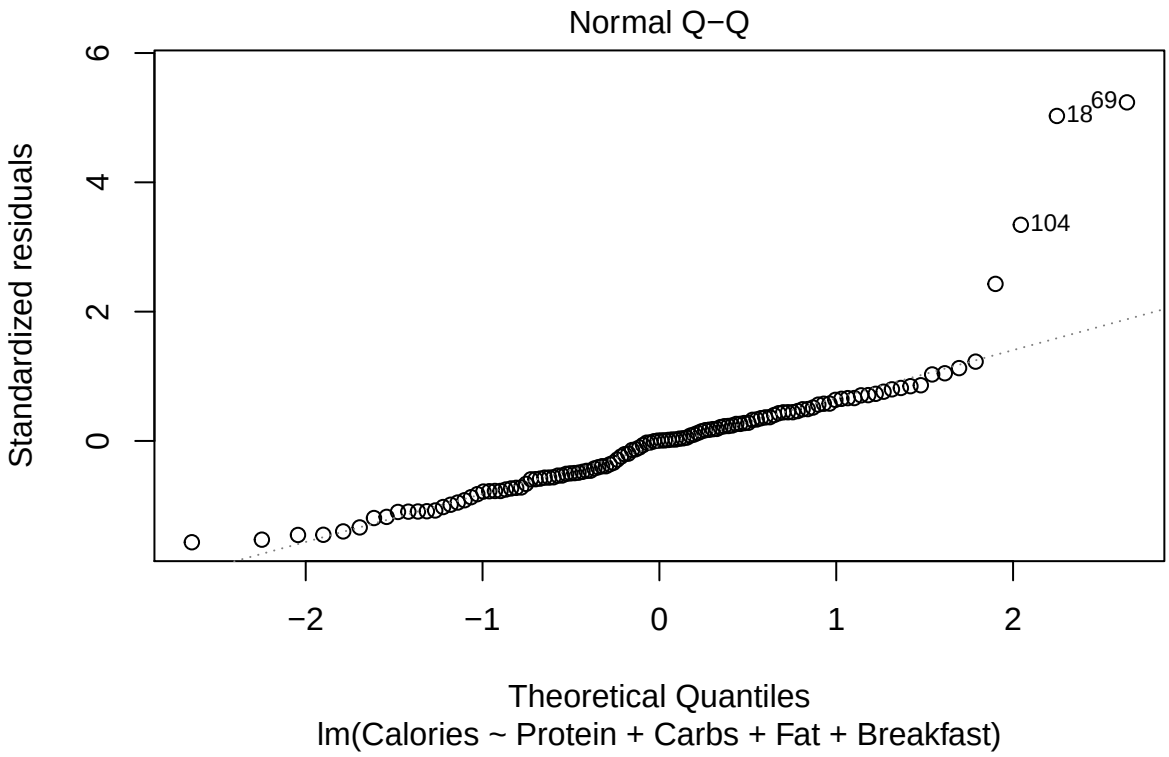
```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.68090    2.00304  -0.340  0.73452
## Protein       3.97538    0.09770  40.691  < 2e-16 ***
## Carbs         3.89569    0.04514  86.310  < 2e-16 ***
## Fat           9.12138    0.08876 102.770  < 2e-16 ***
## BreakfastYes  5.66722    1.70353   3.327  0.00117 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.401 on 117 degrees of freedom
## Multiple R-squared:  0.9987, Adjusted R-squared:  0.9987
## F-statistic: 2.28e+04 on 4 and 117 DF,  p-value: < 2.2e-16
```

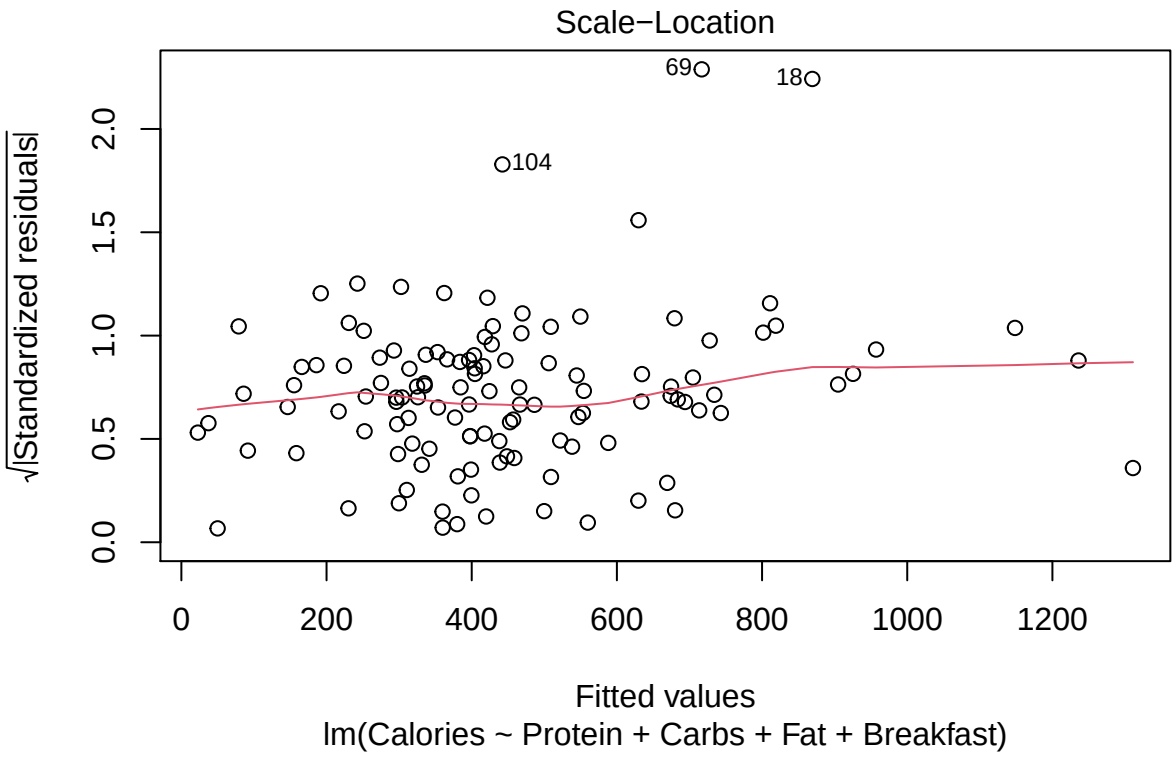
Although the values are slightly different, the points that were made in the regression paragraph for the linear model remain the same. As long as the F statistic is high and the p-value reduced and less than 0.05, the linear model is considered to be well adjusted to the results.

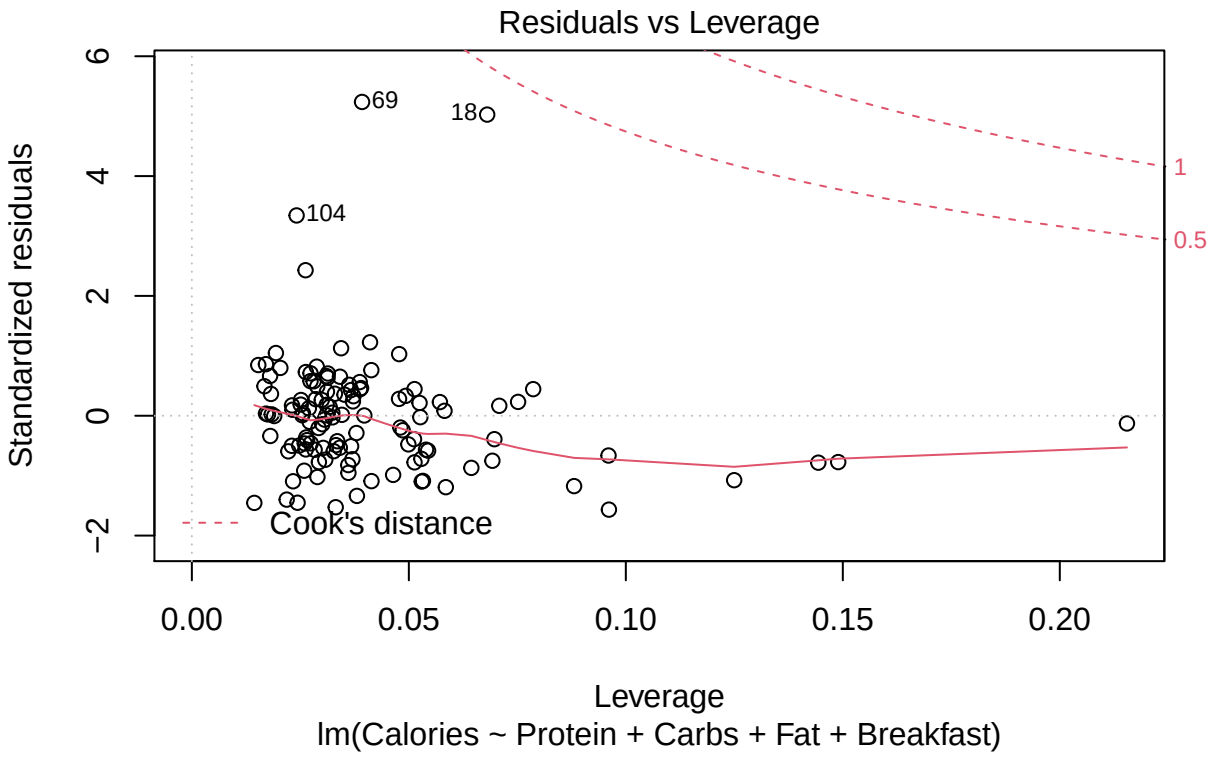
Linear model plots

```
plot(step_b)
```




Normal Q−Q
Standardized residuals
0
2
4
6
18
69
104
−2
−1
0
1
2
Theoretical Quantiles
lm(Calories ~ Protein + Carbs + Fat + Breakfast)


Scale−Location
69
18
104
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
200
400
600
800
1000
1200
Fitted values
lm(Calories ~ Protein + Carbs + Fat + Breakfast)



The results presented here can be summarized as follows:

-In the residuals vs fitted graph, the residual values follow an approximately homogeneous continuous variance by the red line showing 3 outliers represented by the dataset data at positions 104, 69 and 18.

-The Normal Q-Q plot allows us to check if the data follow a normal distribution. If the linear line passes through the points relative to the quantiles of the residuals as a function of the quantiles of a standard normal distribution, it can be stated that the data follow a normal distribution. In this case, this is not the case due to the outliers referred to in the previous point.

-In this model no leverages were found due to the fact that there are no points located in the area between the two red lines in the plot of residuals as a function of leverage. Values that are outliers however are influential observations due to having extreme values of y.

Normality and zero mean

Another way to confirm whether or not the statistic follows a normal distribution is through a shapiro test.

```
shapiro.test(step_b$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  step_b$residuals
## W = 0.80344, p-value = 1.75e-11
```

To be considered as normally distributed, the residuals should give a w-value equal to or very close to 1. If the p-value is smaller than the chosen alpha level, the null hypothesis that the data are normally distributed is rejected. For both conditions the statistical test proves that there is no normal distribution in this dataset.

Cook distance check

It was seen if there were points of cook distance above 1 unit. The result below returned a null vector indicating the inexistence of these points.

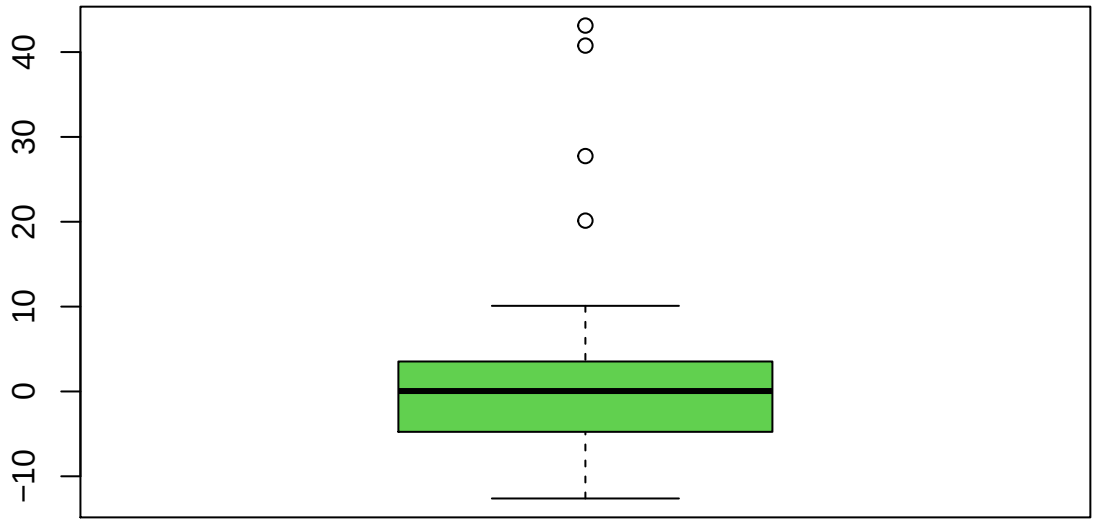
```
cook = cooks.distance(step_b)
pontInf=which(cook>1)  # Para ver os pontos com distância de cook superior a 1 unidade
pontInf
```

```
## named integer(0)
```

Residual outliers

The visual verification of the residuals considered as outliers is represented in the following boxplot.

```
boxplot(step_b$residuals,col=3)
```



4 outlier points were found. To obtain more information about these points, the boxplot.stats function was used.

```
boxplot.stats(step_b$residuals)
```

```
## $stats
##          115           15           49           50           69
## -12.61376541  -4.76067758   0.05017032   3.53244990  10.08964772
##
## $n
## [1] 122
##
```

```
## $conf
## [1] -1.136132  1.236473
##
## $out
##       18       69      103      104
## 40.76405 43.12442 20.13396 27.74330
```

It is possible to extract from this function that the data in the dataset that originate these residues are in positions 18 , 69 , 103 and 104.

Linear model without outliers

In this part, it was decided to remove the residual outliers resulting from the linear model to verify if the model would change in any way. In order not to change the results obtained previously, it was necessary to create a copy of the original dataset, only after this copy was created was the data removed in the positions that originated outlier residues.

```
burger2 <- burger
burger3 <- burger2[-c(18,69,103,104), ]
```

The stepwise algorithm was again used to calculate the AIC value assuming a "backward" direction.

```
yhat2 <- lm(Calories ~ Protein + Carbs + Fat + Meat + Breakfast,data = burger3)
step_b2 <-step(yhat2,direction = "backward")
```

```
## Start:  AIC=397.82
## Calories ~ Protein + Carbs + Fat + Meat + Breakfast
##
##             Df Sum of Sq    RSS     AIC
## - Meat       1         2   3106  395.91
## <none>                     3104  397.82
## - Breakfast  1       421   3524  410.82
## - Protein    1     91912  95016  799.55
## - Carbs      1    516048 519152  999.93
## - Fat        1    703930 707034 1036.38
##
## Step:  AIC=395.91
## Calories ~ Protein + Carbs + Fat + Breakfast
##
##             Df Sum of Sq    RSS     AIC
## <none>                     3106  395.91
## - Breakfast  1       428   3534  409.15
## - Protein    1    114180 117286  822.40
## - Carbs      1    526701 529807 1000.33
## - Fat        1    704221 707327 1034.43
```

```
summary(step_b2)
```

```
##
## Call:
## lm(formula = Calories ~ Protein + Carbs + Fat + Breakfast, data = burger3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.9118  -3.1210   0.4599   3.6347  12.0607
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.09639    1.25436   0.077 0.938883
## Protein       3.95561    0.06137  64.451  < 2e-16 ***
## Carbs         3.92047    0.02832 138.426  < 2e-16 ***
## Fat           9.03714    0.05646 160.063  < 2e-16 ***
## BreakfastYes  4.31518    1.09310   3.948 0.000138 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.243 on 113 degrees of freedom
## Multiple R-squared:  0.9995, Adjusted R-squared:  0.9995
## F-statistic: 5.538e+04 on 4 and 113 DF,  p-value: < 2.2e-16
```

As what happened before, this algorithm assumed the meat variable as the most disposable, however the most interesting result is the fact that the AIC value dropped significantly (from 524.21 in the model with outliers to 395.91 without them).

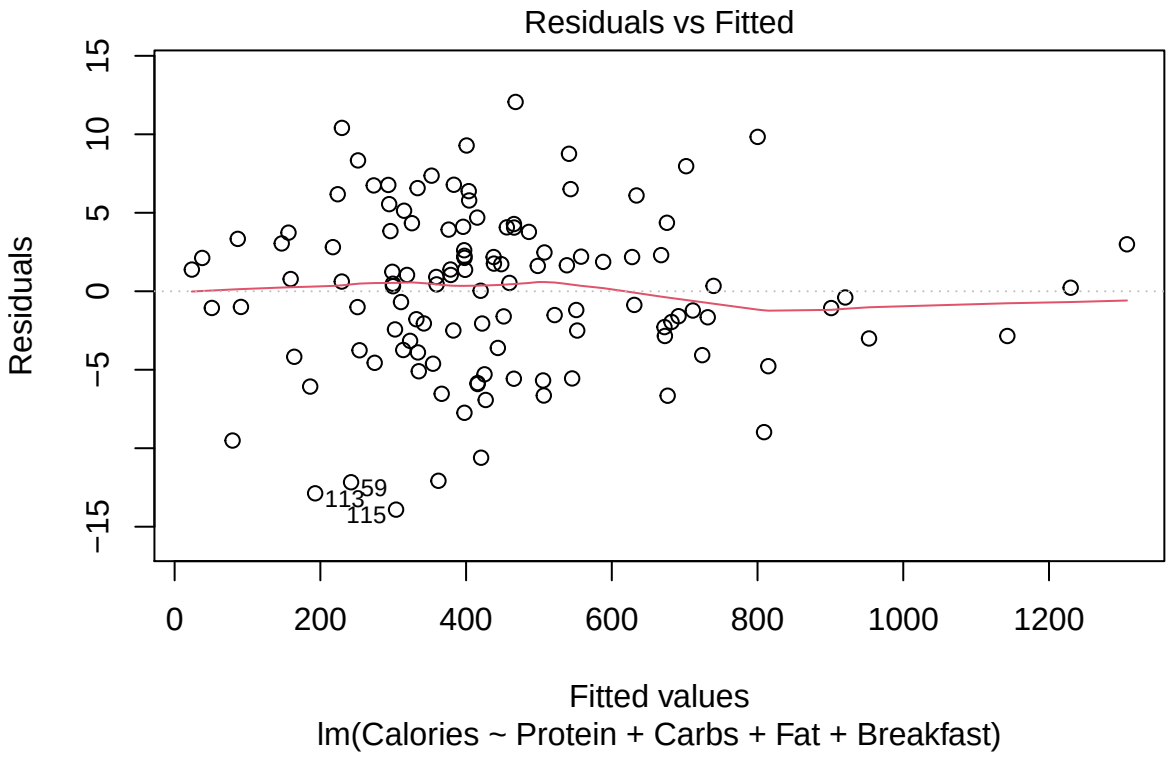
Removing the meat variable also resulted in a slightly greater decrease in the AIC value compared to the original model, in this case the model without outliers (initially 397.82 then 395.91) dropped 1.91 points while in the linear model with outliers the decrease was only 0 ,16 points. This result is interesting for the analysis because the foods considered as outliers had meat in their constitution.

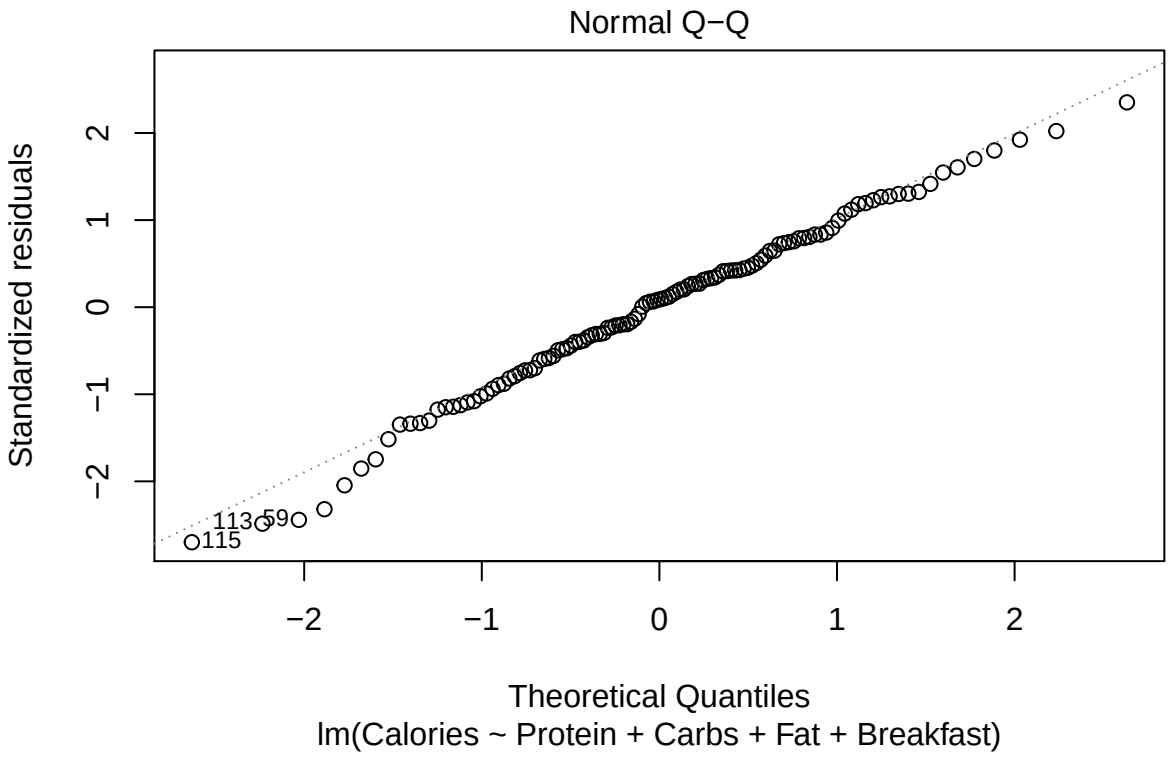
Finally, it can be seen that there was a large increase in the value of the F statistic. The model with outliers had a value of 2.28e+04 while the adjusted model without residual outliers resulted in a value of 5.538e+04, more than twice the original value.

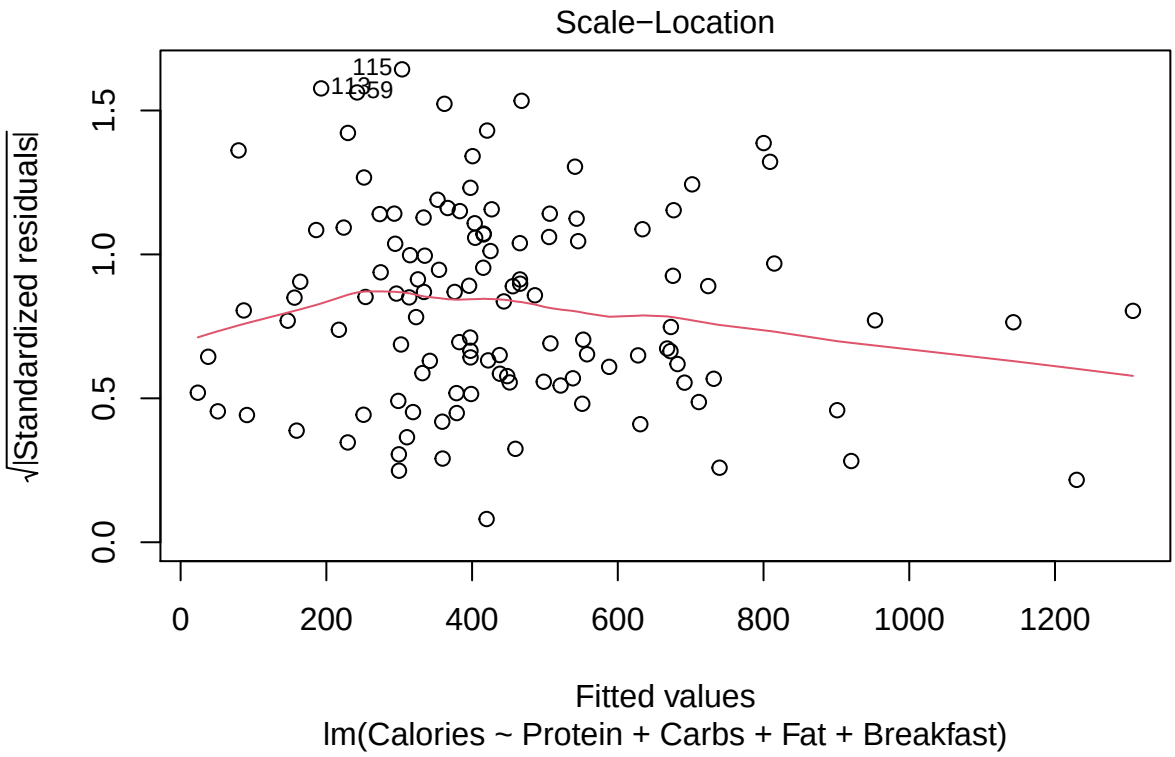
Linear model plots without outliers

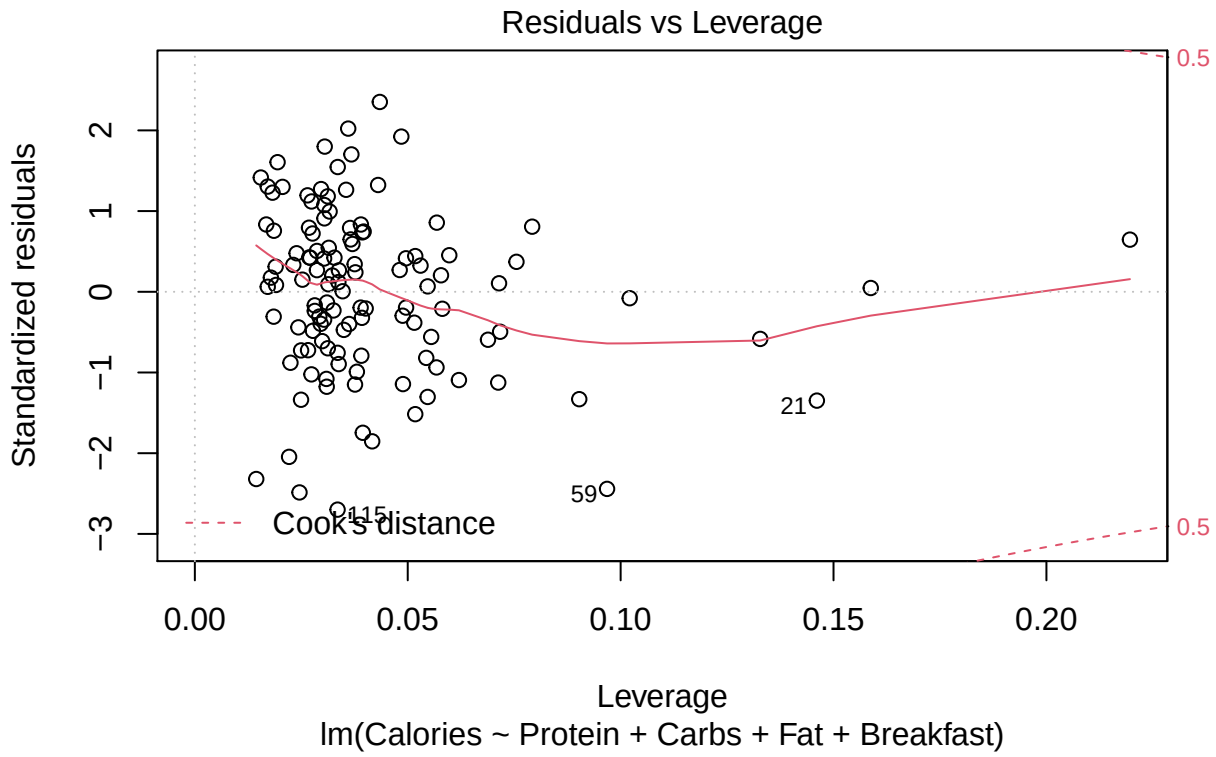
The main difference that removing these residuals should have on the model is to make it more normally distributed.

```
plot(step_b2)
```

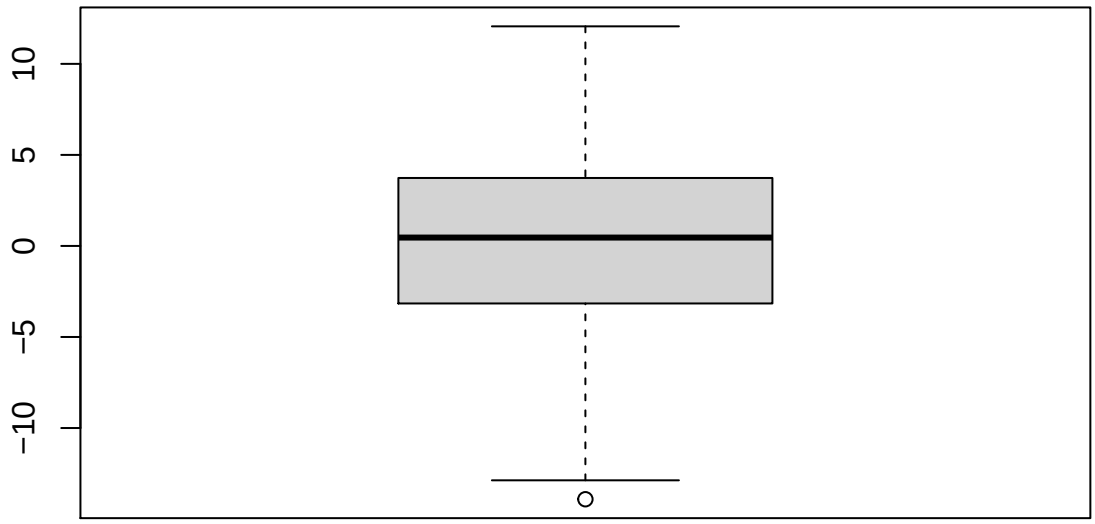

Residuals vs Fitted
Residuals
15
10
5
0
−5
−15
0
200
400
600
800
1000
1200
59
113
115
Fitted values
lm(Calories ~ Protein + Carbs + Fat + Breakfast)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Calories ~ Protein + Carbs + Fat + Breakfast)
113
59
115
−2
−1
0
1
2


Scale−Location
115
113
59
√|Standardized residuals|
0.0
0.5
1.0
1.5
0
200
400
600
800
1000
1200
Fitted values
lm(Calories ~ Protein + Carbs + Fat + Breakfast)




```
boxplot(residuals(step_b2))
```



It can be seen in the quantile quantile graph that the line obtained is closer to a y=x line and most of the points are located on top of the line, however there are 3 new points that compromise normality relative to the data in positions 59, 113 and 115 .

The homogeneity of the variance of the points, as expected, was not compromised. No new points considered as leverage were obtained either.

The model originated a new outlier residue that can be seen in the boxplot obtained. One could try to remove this point to see how the model changes, but at this stage of the work there is already a fairly satisfactory model for the selected parameters.

Normality and zero mean

Using a shapiro test again, a W value close to 1.0 and a p-value greater than 0.5 are obtained, which confirms that the model now has a normal distribution.

```
shapiro.test(step_b2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  step_b2$residuals
## W = 0.99035, p-value = 0.5779
```

Cook distance check

Another null vector was obtained again, so there are no points.

```r
cook2 = cooks.distance(step_b2)
pontInf2=which(cook>1)  # Para ver os pontos com distância de cook superior a 1 unidade
pontInf2
```

```
## named integer(0)
```

Conclusions

With this analysis, it was possible to see that macronutrients are essential for calculating the calorie variable. The analysis confirms that fat has the greatest influence on the value of the caloric content of a food, which goes against the fact that this macronutrient is the most energetic.

It appears that the fact that the food contains meat or is part of breakfast menus did not have as much influence on the caloric content as expected, in the case of meat the relationship with the dependent variable could be discarded without compromising with the feasibility of the linear model.

Finally, it was curious to see that the estimated values for each macronutrient were very close to the actual values. For each macronutrient, the following calculations were obtained in the final linear model without residual outliers:

- Protein had an estimated value in the final linear model equal to 3.95 (actual value = 4 calories/gram)
- Hydrates had an estimated value in the final linear model equal to 3.92 (actual value = 4 calories/gram)
- Fat had an estimated value in the final linear model equal to 9.04, (actual value = 9 calories/gram)

This result together with the values obtained for the F statistic, p-value and the AIC value confirm that the statistical model is very well adjusted to confirm the initially suggested hypothesis.