Abstract

In our times, with the rapid development of medicine and the revolution of the food industry, the public has a demand to know more about nutritional facts more than ever before. In this paper, I will explore a table of natural occurring nutritional values in foods, examining the relationship that arises between certain nutrients such as energy content, fat, carbohydrates and sugar. It is well known that carbohydrates and fat are dense in energy, and this project aspires to verify that relationship. Furthermore, a potential correlation between the contents of B vitamins in foods will be explored. The tools being used are Excel and R, with which I cleared, transformed, investigated and modelled the data, and mainly fitted linear regression models to it. Indeed, the results will indicate that carbohydrate, energy and fat content of food highly influences the total energy content it provides, and an underlying relationship between B vitamins will be investigated.

Introduction

The data set I chose belongs to the category of health data sets, and it exhibits the nutritional information of a variety of foods. It contains 8,789 food entries, together with the mean amount of each nutrient they contain for each 100g of food [4]. The nutritional variables it contains are mainly macronutrients, micronutrients, vitamins and minerals. Together with the data set, a user manual is available that can help the reader of the data understand the scientific and statistical processes that were enforced for the dataset to be published. The data appearing in this dataset is a composition of data from the *Agriculture Handbook 8* of the US Department of Agriculture together with its supplements. The data of this book is a synthesis of information from scientific publications that are mentioned in the sources of the book and the manual, and from data of the food industry, other government agencies, and research conducted under contracts initiated by USDA's Agricultural Research Service (ARS) [4].

The dataset initially contained 50 variables, which included the water content of each food, ash, vitamins, minerals, macronutrients and micronutrients.

I decided to focus my analysis on macronutrients and vitamins only, and this happened for the following reasons: First of all, the data had numerous missing values in the variables that I decided to not include. These values cannot be substituted with the means of the rest of the data, since food can be very unpredictable in its nutrient content and it does not follow any regular distribution, especially since different categories of foods present different values that would be considered as extremes in another category. The second reason is interest. Generally, a person is more likely to be interested in the vitamin or macronutrient content of food, and not consider the values of the rest of the nutrients. In fact, there were a lot of nutrients that I had never heard of before as someone who does not specialise in diet, and it was not easy to understand their role even after researching about them.

More specifically, my analysis will consist of two parts. The first part is exploring the relationship between fat, carbohydrates and sugar with the energy content of food. Knowing that these three nutrients are foundational energy sources in our diet, I wanted to explore how do they contribute to the total energy content, and to confirm that this relationship is indeed real. The second part is a question that occurred to me while I was exploring the correlations between nutrients. It seemed like there was a medium positive correlation between the amounts of vitamins of the B complex that I was focusing on.

Methods

When I first downloaded the dataset, it was in Excel format, and therefore I decided to do some data clearing in Excel first, since I was more familiar with it rather than R. I decided to focus on the certain variables that I mentioned and therefore I deleted the other columns that were not of interest. After deleting the columns, I proceed to delete the remaining entries that contained gaps. Then, I imported my initial data set in R, called “abbrev” and my final data set called “abbrev4”.

Following this, I decided to start properly exploring the numerical properties of the data set that I had isolated. Starting with the summary() function in R, I could get a better understanding of the set, and obtain a feeling of its statistical properties, like the mean of each variable, variance, minimum and maximum values and so forth. Moving forward, I decided to explore the possible correlation between certain variables. Using the cor() function in R, I gathered all the correlations in a table, and tried to find variables that would have a positive or negative correlation that has an absolute value larger than 0.5, as this can indicate some strong relationship. Based on where I saw correlation, I isolated these cases and decided to ask certain questions, together with my own understanding of nutrients and the basic knowledge I had.

Question 1: Is the energy content of foods related to the total fat content?

For this, I created a new variable called ‘Total Fat’, which is the sum of saturated and unsaturated fats for each food. I set the total fat as independent variable, and energy as the dependent variable. To decide to use this method, I used the ggplot2 command to make a scatter plot of these variables, together with the built-in function that provides a linear fit on the plot. It looked encouraging, and therefore I did the regression, found the regression parameters and coefficients, and used the R function plot() to see the basic graphs that would help validating the assumptions of the model.

Then I decided to further explore the data by isolating the first word of the food description, which denotes the basic ingredient of the food. I did this so I could explore food categories more in depth. The categories I worked with were cheeses, meats and fruits (including fruit juice). I manually looked at the data set, knowing by the user manual that it was ordered depending on food categories, and isolated these food entries in distinct data frames for each of the three categories. I used similar methods with before, looking at correlations in these foods and performing regression.

For meat and cheese, I fit regression models with energy and total fat, and the results seemed clearer than the general category of all foods. For fruit, since I know that their fat content is generally low, I became curious about the relationship between the energy content in food and its sugar and carbohydrate content. Again, I fitted a linear regression model for both cases, and examined a third case with multiple variable linear regression. This motivated me to form another idea, that the initial model for energy in foods would be improved if I did multiple linear regression with total fat and carbohydrates. Indeed, the model was better, and this is logical because energy content in foods stems from multiple nutrients, rather than just fat.

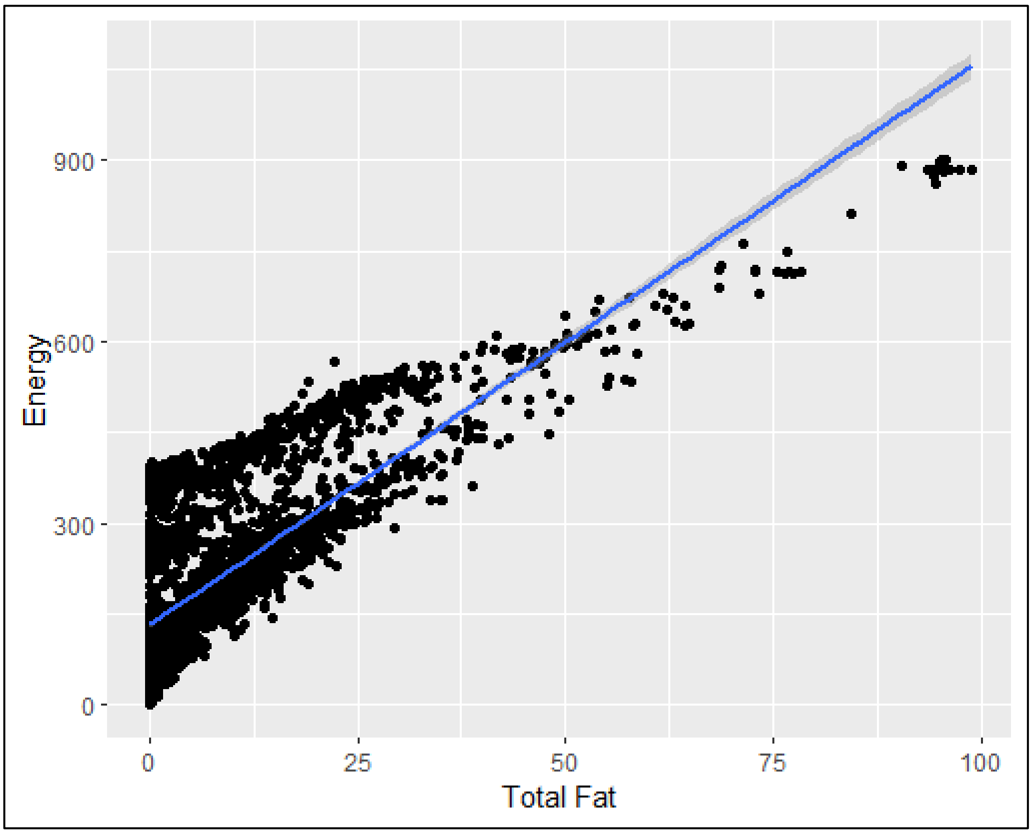
After exploring all these, two more questions occurred to me that were most interesting. In my initial correlation table, I had observed that there is a somehow medium to strong positive correlation between amounts of B vitamins, approximately 0.6 in most cases. This made me curious. Why do these vitamins occur in foods together? While doing research in bibliography, I realised that, although it was not mentioned explicitly that they appear together, they have a similar role in how they work in organisms, either animals or plants, and they interact with each other. Thus it makes sense that they will be found together in foods, especially not heavily processed foods [1][2].

I chose to fit a multiple linear regression to b vitamins, taking one of them as dependent variable and the rest as independent variables, but the model was not very good. After looking at the assumptions, I decided to eliminate outliers that significantly affected the model, and I removed mainly them because I traced down which foods were outliers. It was always a fortified product, so it did not help with my questions that have to do with the natural occurrence of nutrients in food.

Results

Question 1: Is there any relationship between energy content in foods and their total fat content?

After looking at the correlations table, I noticed that the correlation between energy vs saturated, monounsaturated and polyunsaturated fat was approximately 0.64, 0.69 and 0.6 respectively. This motivated me to investigate if there is a linear relationship between energy and total fat. This is the quick scatter plot of the two variables with the ggplot2 fitted linear model, which indicates some linearity:

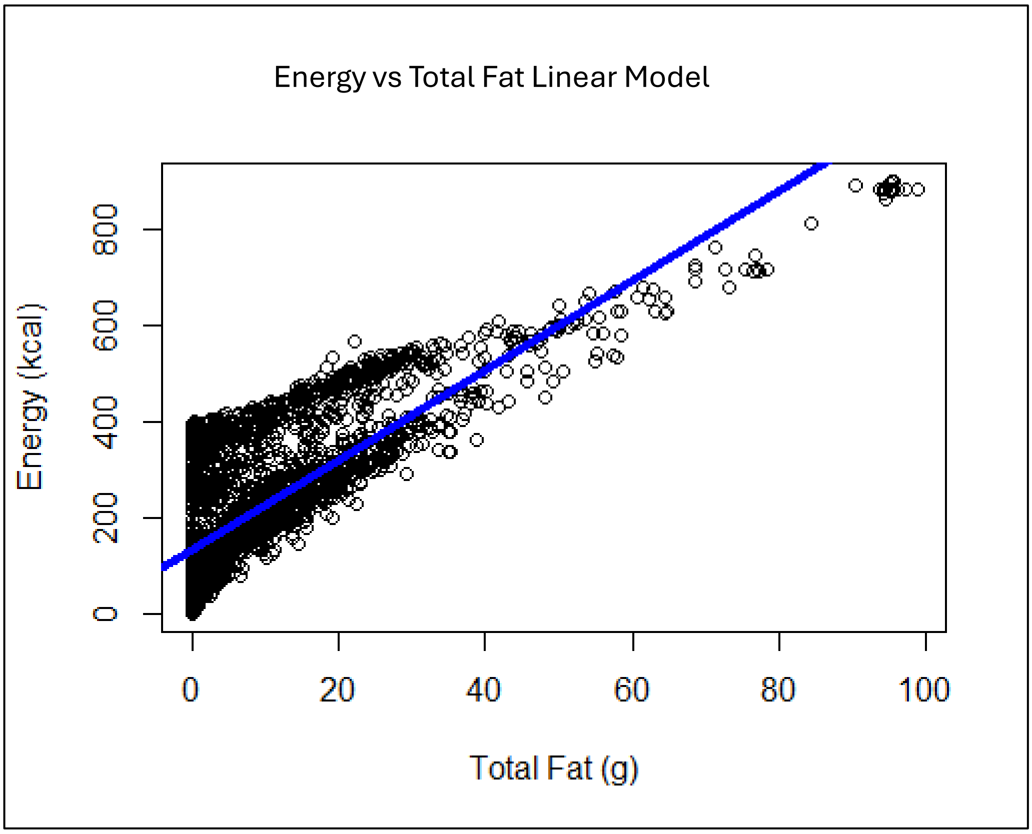


Here are the results of the first model I fitted to the data:

```
Residuals:
    Min      1Q  Median      3Q     Max 
-171.02  -73.57  -29.47   34.69  265.02 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 132.7304     1.9692   67.40   <2e-16 ***
total_fat_g   9.3365     0.1182   78.96   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101.2 on 3685 degrees of freedom
Multiple R-squared:  0.6285,	Adjusted R-squared:  0.6284 
F-statistic:  6235 on 1 and 3685 DF,  p-value: < 2.2e-16
```

The R squared is 0.6285, which is not very strong, but can indicate that the model might highlight a potential relationship. The p value is lower than 0.05, and therefore there is statistical significance which suggests that linearity is a possibility.

This is the linear regression line:

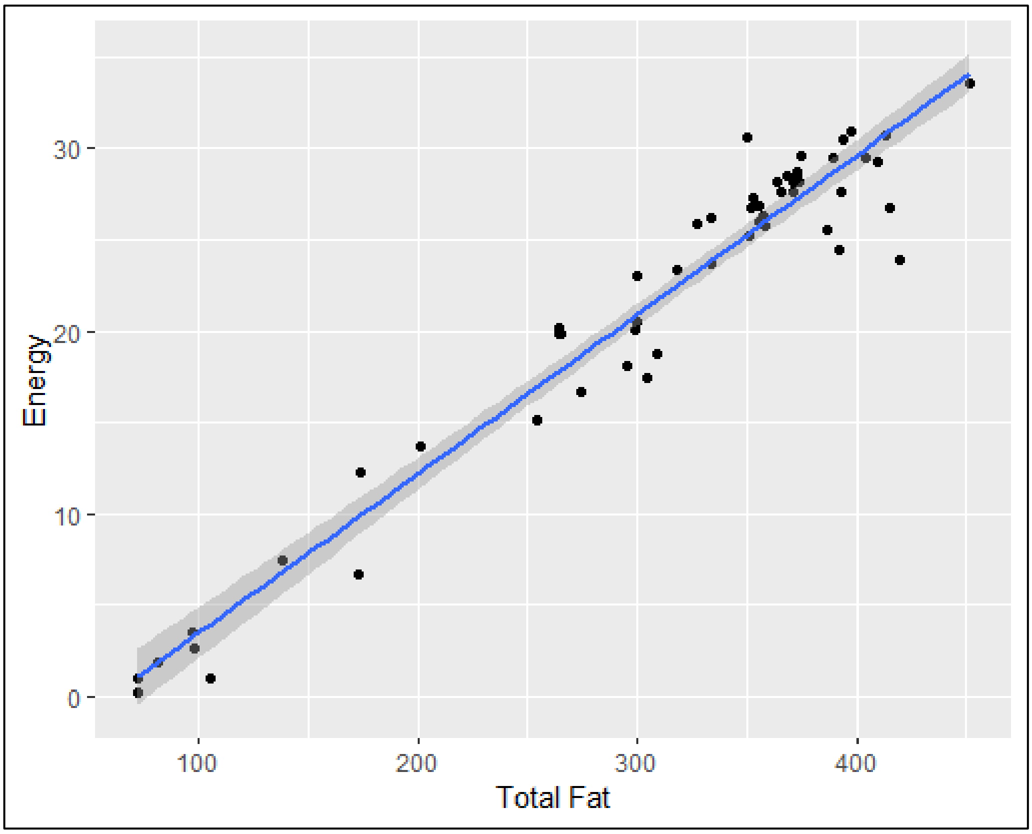


Going more in depth:

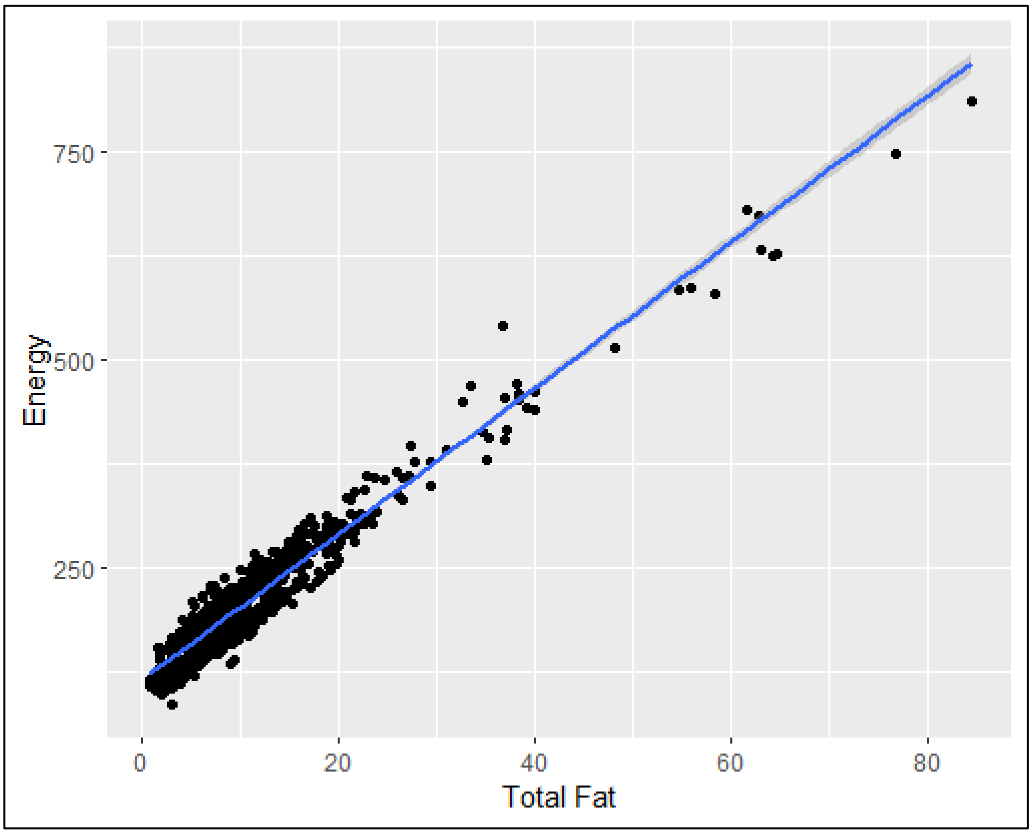
Question 2: Is there a stronger relationship between energy and fat in specific food categories, such as cheese and meat?

Knowing that cheese and meat are generally higher in fat on average compared to other foods, I looked at two ggplot2 plots with the built in linear model fitted to decide if there might be a linear relationship between these two nutritional variables in these food categories.

This is the plot for cheese:



The plot for meat:



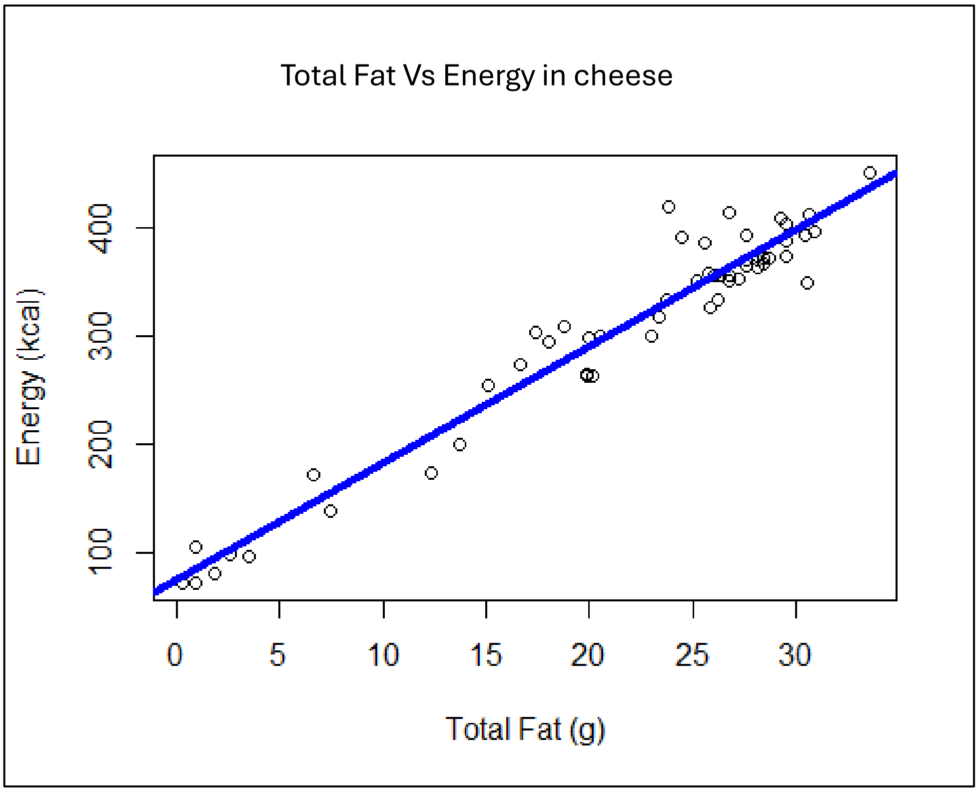
Both plots indicate linearity, so I proceeded with the linear regression. These are the results:

Regression for cheese:

```
Residuals:
    Min      1Q  Median      3Q     Max 
-55.501 -14.893  -5.215  13.310  87.151 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   74.447      8.767   8.492 2.14e-11 ***
total_fat_g   10.818      0.373  29.001  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.85 on 52 degrees of freedom
Multiple R-squared:  0.9418,	Adjusted R-squared:  0.9407 
F-statistic: 841.1 on 1 and 52 DF,  p-value: < 2.2e-16
```

This time, R squared indicates a much stronger relationship than before, being larger than 0.9, which might suggest there is a strong linear link between energy and fat in cheese. This makes sense, because cheese has a great fat content and not such a great content in carbohydrates so its calorie content must have a strong dependence on fat. The p value is again less than 0.5 and so this is a sign of existence of this linear relationship.

Plot for the linear model for cheese:

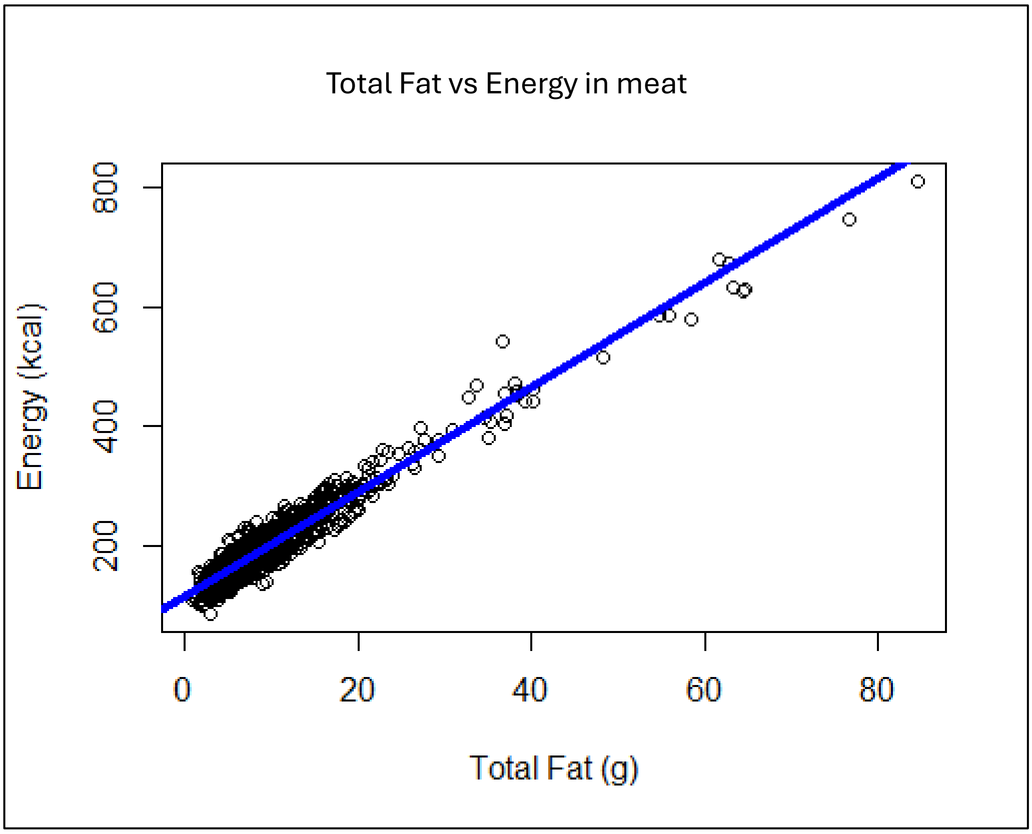


Regression for meat:

```
Residuals:
    Min      1Q  Median      3Q     Max 
-59.940 -17.834   2.666  17.137 103.499 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 114.46980    1.15010   99.53   <2e-16 ***
total_fat_g   8.77634    0.08343  105.19   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.91 on 832 degrees of freedom
Multiple R-squared:  0.9301,	Adjusted R-squared:   0.93 
F-statistic: 1.107e+04 on 1 and 832 DF,  p-value: < 2.2e-16
```

For meat too, we have a very high value of R squared and a value of p that is less than 0.5, and therefore it seems that there is a significant linear relationship between energy and fat for this category.

Plot for linear model in meat:

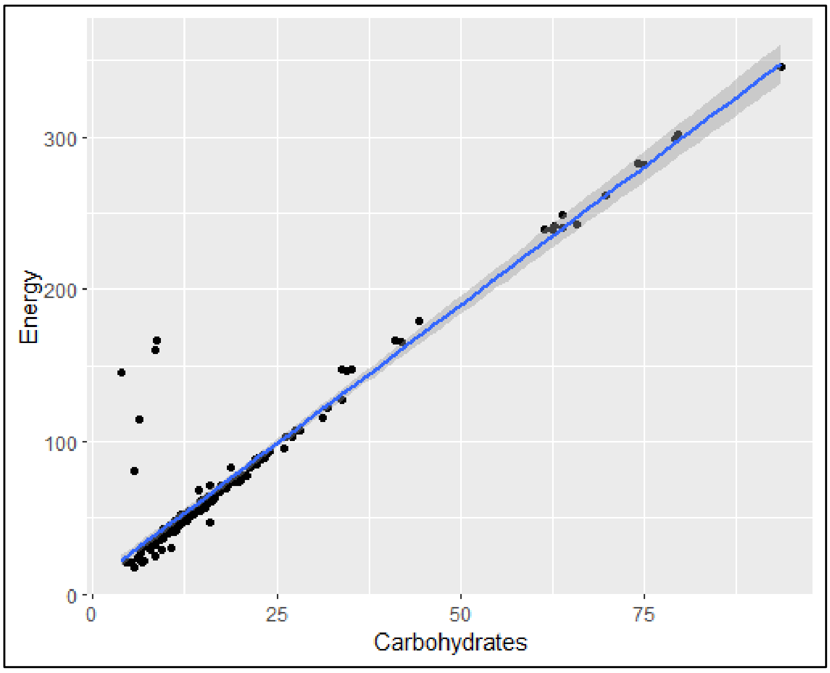


Question 3: Is there a relationship between the energy content of fruit and fruit juice, and the amount of sugar or carbohydrates they contain?

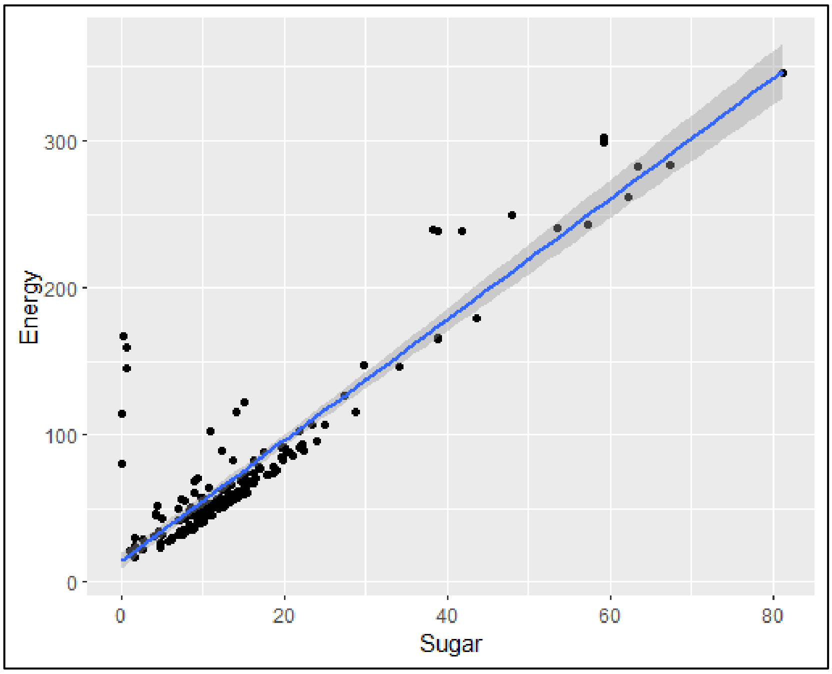
Knowing that fruits and juice generally contain trace amounts of fat, but are rich in sugar and carbohydrates, I decided that this would be a more suitable and interesting question for this category.

This is the initial idea we get from some quick ggplot2 plots:

Energy vs carbohydrates plot:



Energy vs sugar plot:



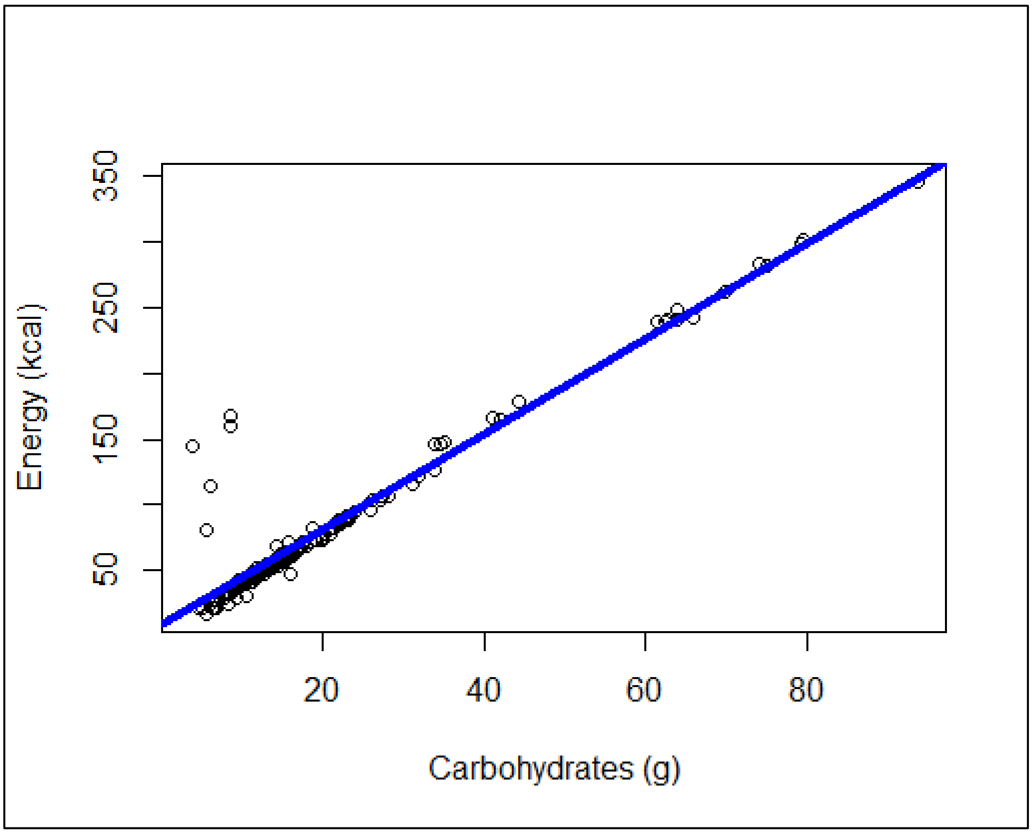
Again, there is an underlying linear relationship.

Proceeding with the energy vs carbohydrate regression:

```
Residuals:
    Min      1Q  Median      3Q     Max
-19.055  -4.860  -3.173  -1.479 127.709

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.87143    2.07483   3.794 0.000201 ***
carbohydrt_g  3.63649    0.08527  42.646  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.11 on 186 degrees of freedom
Multiple R-squared:  0.9072,    Adjusted R-squared:  0.9067
F-statistic:  1819 on 1 and 186 DF,  p-value: < 2.2e-16
```

Both R squared and p value are very encouraging that a linear relationship does exist.

Plot:



Energy and sugar regression:

```
Residuals:
    Min      1Q  Median      3Q     Max 
-16.784 -12.081  -6.944   0.520 151.569 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  14.1993     2.7799   5.108 8.03e-07 ***
sugar_tot_g   4.1049     0.1393  29.475  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.97 on 186 degrees of freedom
Multiple R-squared:  0.8237,	Adjusted R-squared:  0.8227 
F-statistic: 868.8 on 1 and 186 DF,  p-value: < 2.2e-16
```

Good R squared value and p value.

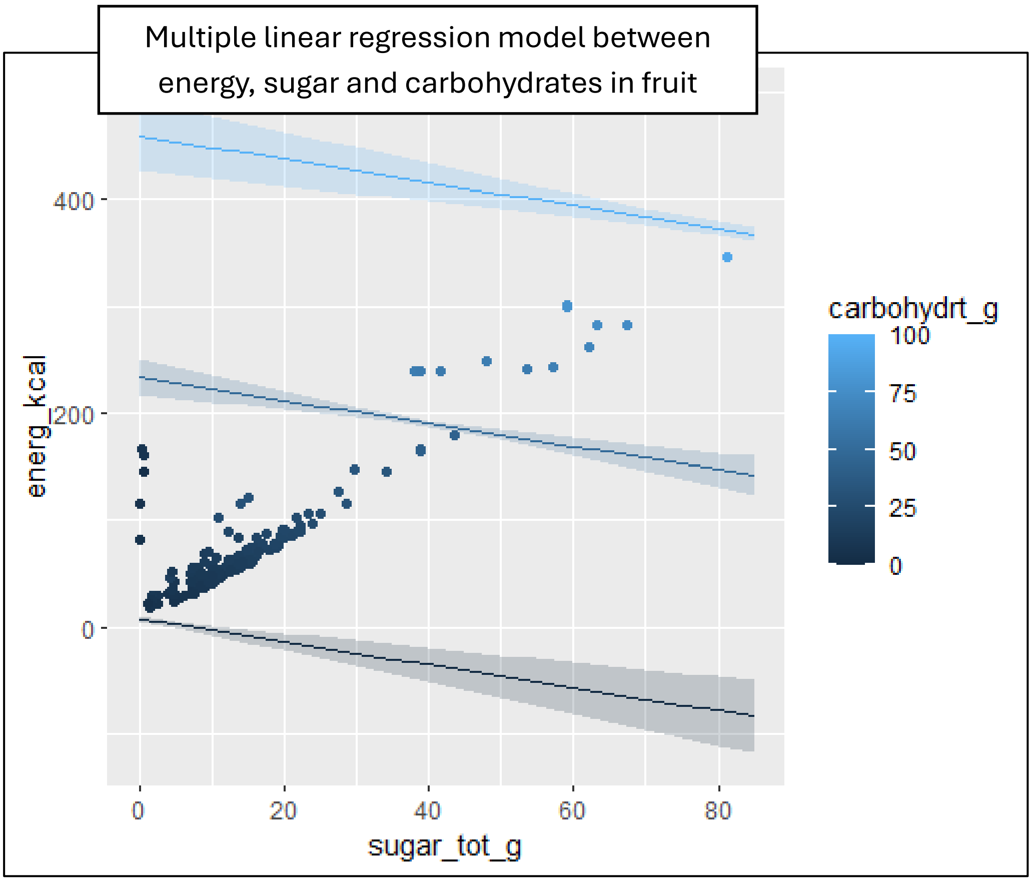
However, I wondered if the models could be improved, and since both nutrients seem to have a strong relationship with the calorie content, I decided to explore this further and fit a multiple linear regression model, that would have both carbohydrates and sugar as independent variables and energy as a dependent variable.

These are the results of the multiple linear regression:

```
Residuals:
    Min      1Q  Median      3Q     Max 
-28.390  -4.445  -2.640  -0.378 120.691 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   7.5941     2.0429   3.717 0.000267 ***
sugar_tot_g  -1.0789     0.3984  -2.708 0.007399 ** 
carbohydrt_g  4.5184     0.3363  13.437  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.81 on 185 degrees of freedom
Multiple R-squared:  0.9108,	Adjusted R-squared:  0.9098 
F-statistic:   944 on 2 and 185 DF,  p-value: < 2.2e-16
```

The above gives very positive looking values of R squared and p that indicate a strong linear relationship. However, in this linear model there is a negative linear relationship with sugar, despite the fact that sugar can contribute to total calories. This can happen because the carbohydrate content of a food contains the sugar content in this dataset, and consequently we do not gain much by including both carbohydrates and sugars to a regression model.

Visualisation of the model:



We see in the plot that although there is positive correlation between carbohydrates and energy content, there is negative correlation between sugar and energy, suggesting that there might be a problem with the model.

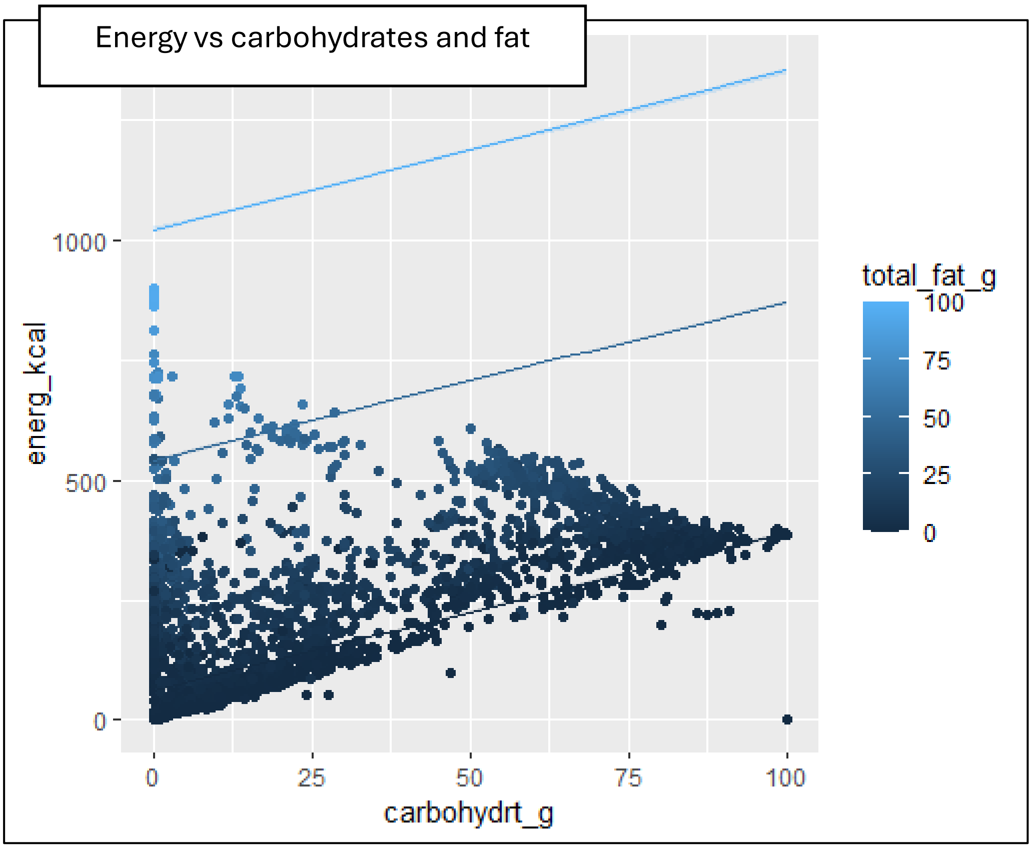
After having these results, I concluded that the initial model could be improved if I added more variables. Energy in food highly depends in its fat, sugar and carbohydrate content as each nutrient contributes several calories to the total. Hence, I fitted a multiple linear regression model to all the foods, with carbohydrates and total fat as independent variables, since the carbohydrate content contains the sugar content.

These were the results:

```
Residuals:
    Min      1Q  Median      3Q     Max
-390.92  -44.43    1.44   33.54  301.62

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  59.57371    1.15369   51.64   <2e-16 ***
carbohydrt_g  3.31345    0.02990  110.81   <2e-16 ***
total_fat_g   9.63485    0.05687  169.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 48.6 on 3684 degrees of freedom
Multiple R-squared:  0.9143,	Adjusted R-squared:  0.9142
F-statistic: 1.964e+04 on 2 and 3684 DF,  p-value: < 2.2e-16
```

This model is significantly better than the very first model, mainly because of how much higher the R squared is. This means that there probably is a much stronger linear relationship between energy and fat and carbohydrates in food, rather than just between energy and fat.

Visualisation of the model:



There was a final question that I would like to address, with not very clear results however, but it was a surprising find while looking at my data. For this reason, I think it is worth mentioning.

While examining the correlation table for all foods, I encountered a possible relationship between pairs of vitamins that belong to the B complex. This is the table for reference:

	thiamin_mg	riboflavin_mg	niacin_mg	panto_acid_mg	vit_b6_mg	vit_b12_ug
thiamin_mg	1.0000000	0.6266527	0.6463639	0.5117470	0.6005904	0.1715389
riboflavin_mg	0.6266527	1.0000000	0.6998804	0.6445982	0.6905050	0.5064669
niacin_mg	0.6463639	0.6998804	1.0000000	0.6242997	0.8219630	0.3945418
panto_acid_mg	0.5117470	0.6445982	0.6242997	1.0000000	0.5786739	0.3736904
vit_b6_mg	0.6005904	0.6905050	0.8219630	0.5786739	1.0000000	0.3962462
vit_b12_ug	0.1715389	0.5064669	0.3945418	0.3736904	0.3962462	1.0000000

Not being very familiar with a more suitable model for a situation like this, I decided to fit all the B vitamins, except from vitamin B12, which is only contained in animal products [3] and therefore displays lower correlation, in a single multiple linear regression model, keeping vitamin B6 as dependent variable and all the rest as independent. This is the resulting model:

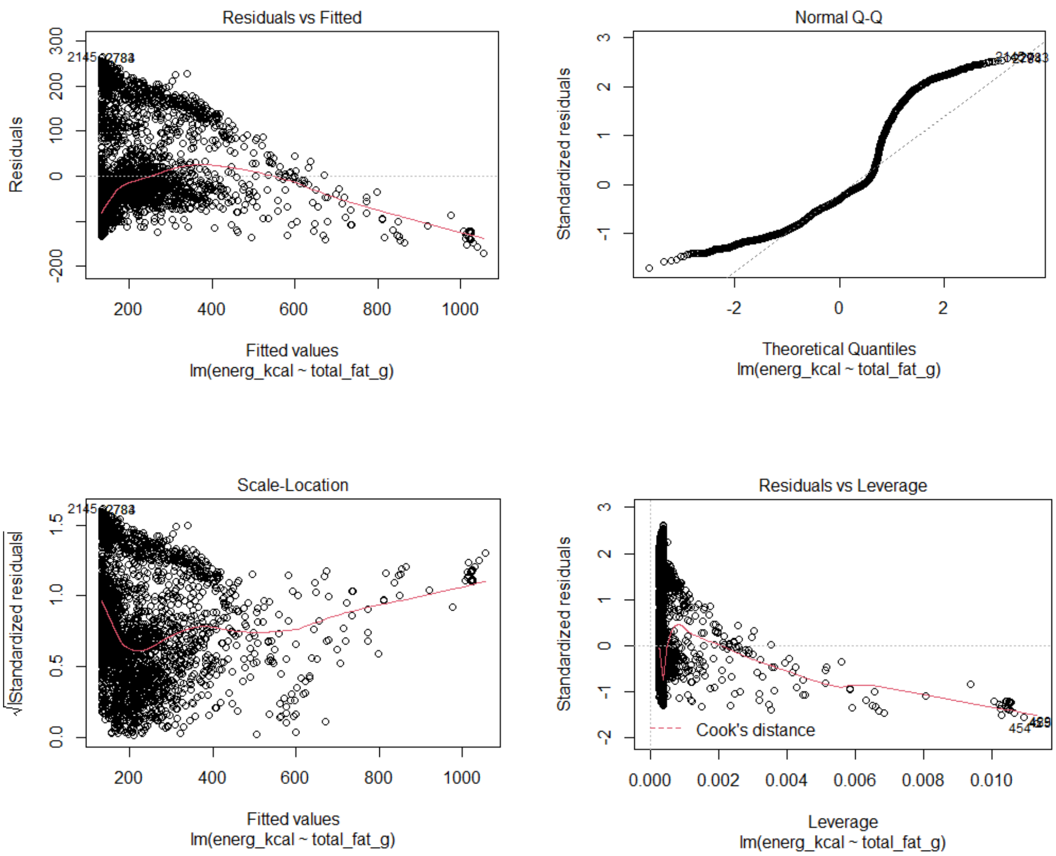
```
Residuals:
    Min      1Q  Median      3Q     Max 
-2.5200 -0.0652 -0.0064  0.0613  5.2829 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.008118   0.005069   1.602   0.1093    
thiamin_mg    0.058714   0.013338   4.402  1.1e-05 ***
riboflavin_mg 0.218480   0.015916  13.727  < 2e-16 ***
niacin_mg     0.059015   0.001323  44.592  < 2e-16 ***
panto_acid_mg 0.008911   0.003525   2.528   0.0115 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2463 on 3643 degrees of freedom
  (36 observations deleted due to missingness)
Multiple R-squared:  0.7039,	Adjusted R-squared:  0.7036 
F-statistic:  2165 on 4 and 3643 DF,  p-value: < 2.2e-16
```

R squared does not indicate a significantly strong model, but not a very weak one either. We could say there is a possibility that those nutrients may be related in some way. In fact, they are, because B vitamins all work together to ensure that metabolism works properly in organisms and the absorption and distribution of energy happens sufficiently, and therefore it makes sense for them to be found in similar foods [1][2]. However, this model is not sufficient to prove this, since the coefficients are small, so they might be overcome by error which is extremely large compared to the scale of the coefficients. Now, the validity of these models will be discussed in the next section.

Discussion

The linear models I have used take as given four different assumptions about the data. These are: independence of the outcomes, that they are normally distributed, that their mean is a linear function of their corresponding independent variable(s), and that they have the same variance. I will be checking the validity of these assumptions by examining the Q-Q, Residuals vs Fitted and Scale-Location plots.

Question 1:



Independence: Generally, foods that contain different ingredients will have an independent energy content. The problem would arise when foods contain similar ingredients, which can contribute to the energy content in a non-independent way. There are foods that contain similar ingredients in the data set and therefore the different energy contents are not fully independent.

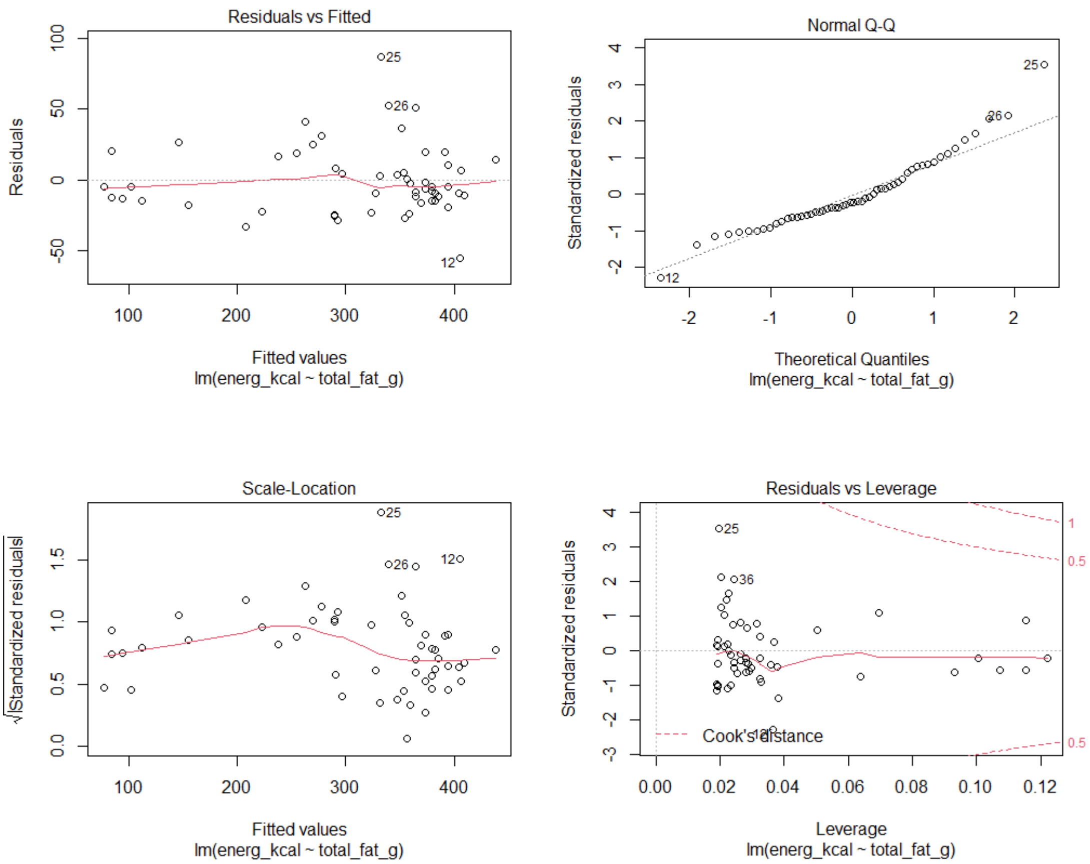
Normality: The Q-Q plot indicates a potential problem with normality. The majority of points deviate significantly from the line, and therefore this might mean that energy and fat are not normally distributed.

Linearity: There is not obvious pattern in the Residuals vs Fitted plot, which is encouraging about linearity between the variables. However, as values get larger the red line seems to be declining, so there could be an underlying issue. Generally, linearity seems to be fine.

Constant Variance: From the Residuals vs Fitted plot, and from the Scale-Location plot, it seems like the assumption of constant variance fails. The red lines are not horizontal, and this indicates that variance increases as the values of the energy increase.

Question 2:

Cheese regression:



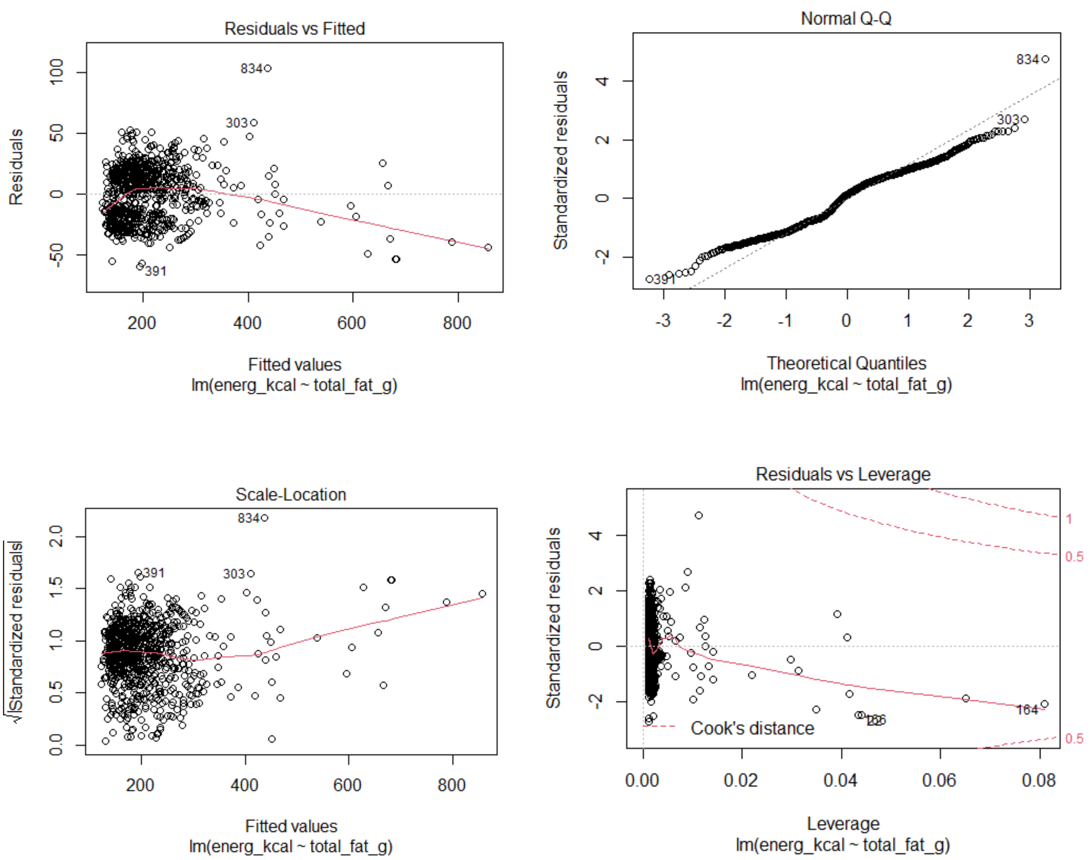
Independence: Independence here can be even worse than the previous question, because it is probable that the base ingredient of a lot of the entries in the cheese category will be the same kind of cheese. Therefore, independence can fail significantly in this model.

Normality: The Q-Q plot is somewhat better for this model. More points seem to be lying near the straight line. Still, it is not very good. Chances are that the data is not normally distributed.

Linearity: The Residuals vs Fitted plot does not indicate any obvious pattern, and the red line is almost vertical, and so we can assume safely that there is a linear relationship between energy in cheese and total fat.

Constant variance: There is not any concerning indication about constant variance violations in the data. Both in the Residuals vs Fitted plot and in the Scale-Location plot there is a curve around the centre of the data, but it is not very large and therefore we can assume that the variance is generally constant among the data.

Meat Regression:



Independence: Similarly to cheese, there should be a lot of foods in this category that contain the same type of meat as a basic ingredient, and this causes dependencies in the data.

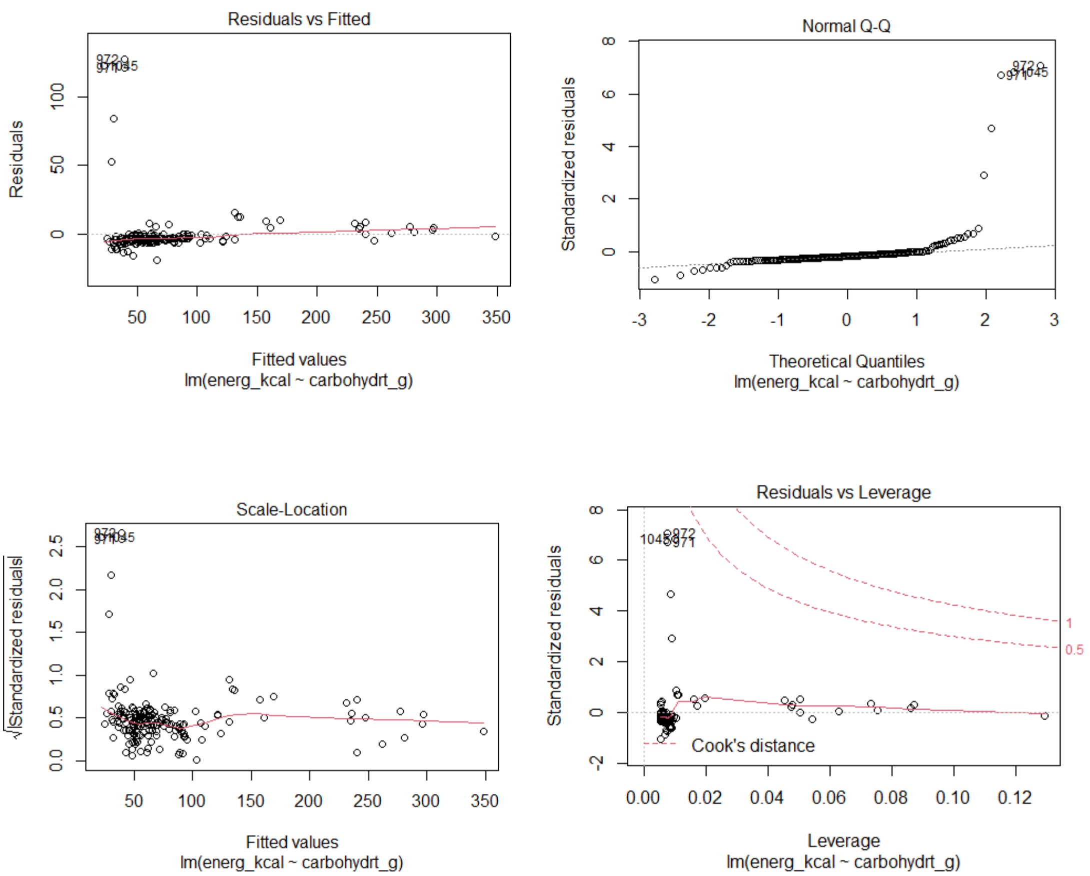
Normality: Generally, the points lie near or on the straight line, but there are some deviations from it in the Q-Q plot, which might suggest that the data is not approaching a normal distribution sufficiently.

Linearity: As values increase, we observe that the red line in the Residuals vs Fitted plot has a very clear negative slope, which is a concerning sign for linearity. Perhaps a linear model is not the best fit for this set of data.

Constant Variance: By looking at the Residuals vs Fitted plot and the Scale-Location plot, we can see that variance is probably not constant and that it increases as the values of the variables increase. Therefore, this assumption might fail for meat.

Question 3:

Regression of energy vs carbohydrates in fruit:



Independence: The data is not independent if the food is based or contains the same type of fruit. More specifically, the set contains fruit and juice, and thus there can be juices that are made from a certain fruit that is also contained as an entry.

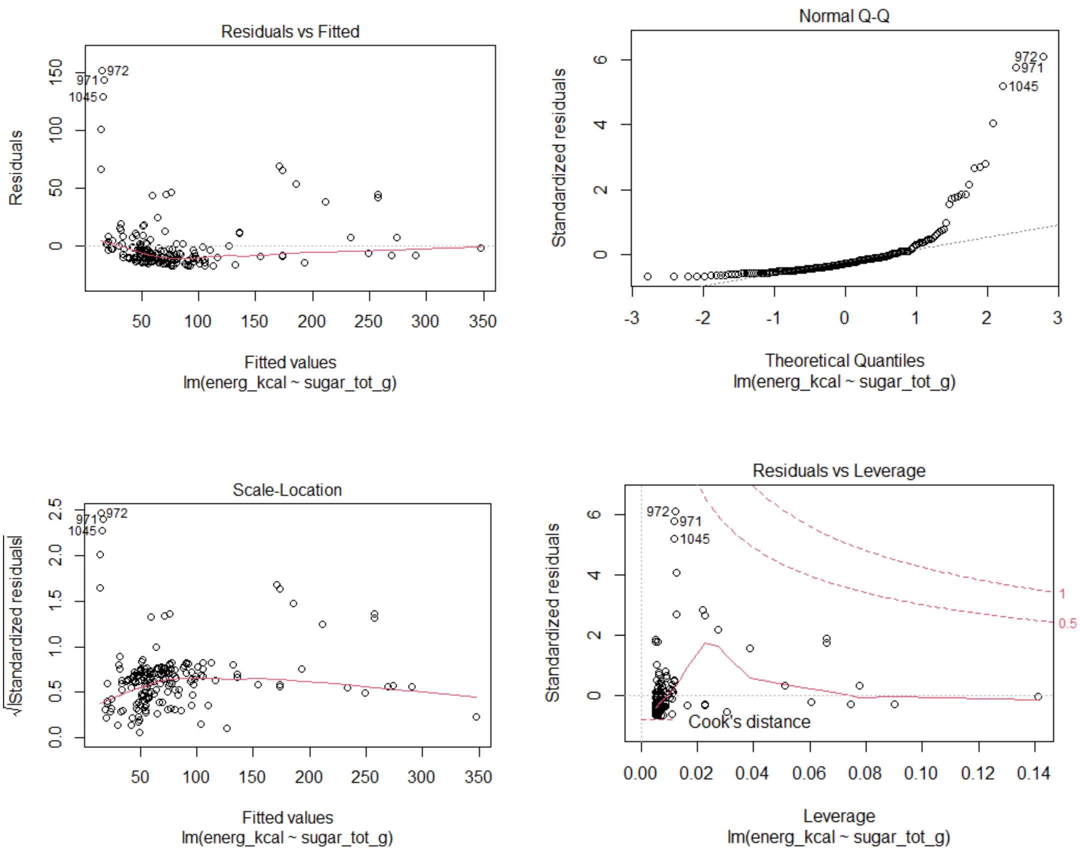
Normality: The Q-Q plot indicates serious issues with normality. There are many points that deviate significantly from the straight line.

Linearity: The red line in the Residuals vs Fitted plot is completely straight and the plot does not indicate any trend, and therefore we can assume that linearity is valid.

Constant variance: Looking at the Residuals vs Fitted and Scale-Location plots, we do not see anything unusual, and the red lines are horizontal for both, meaning that variance is generally constant.

The points might seem to be concentrated and not randomly scattered, but this is because the plots contain some extreme outliers, and hence they are more condensed.

Regression of energy vs sugar in fruit:



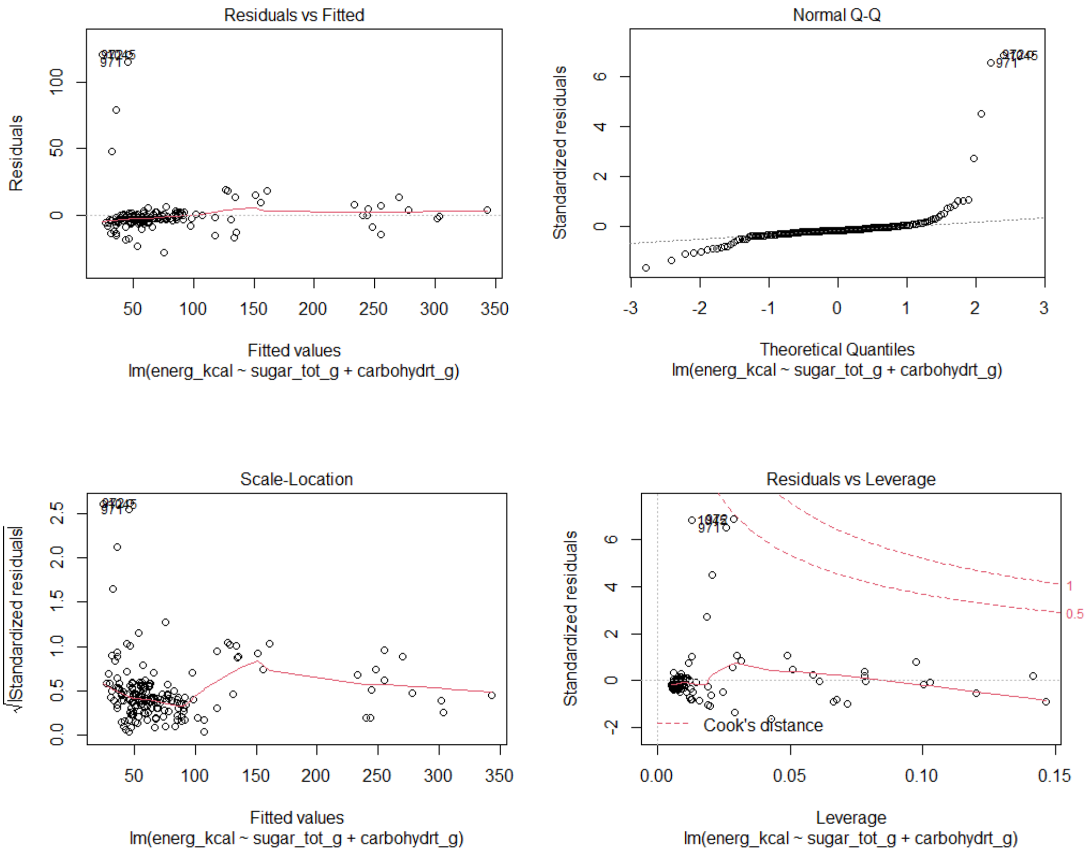
Independence: Similar to the previous case.

Normality: The Q-Q plot suggests issues with normality, as before.

Linearity: The points are generally randomly scattered around the red line in the Residuals vs Fitted plot, and the line is horizontal. It makes sense to assume linearity.

Constant variance: Both in the Residuals vs Fitted and Scale-Location plots the red lines are horizontal and the points are randomly scattered with the exception of some outliers. Variance should be constant generally.

Regression of energy vs sugar in fruit:



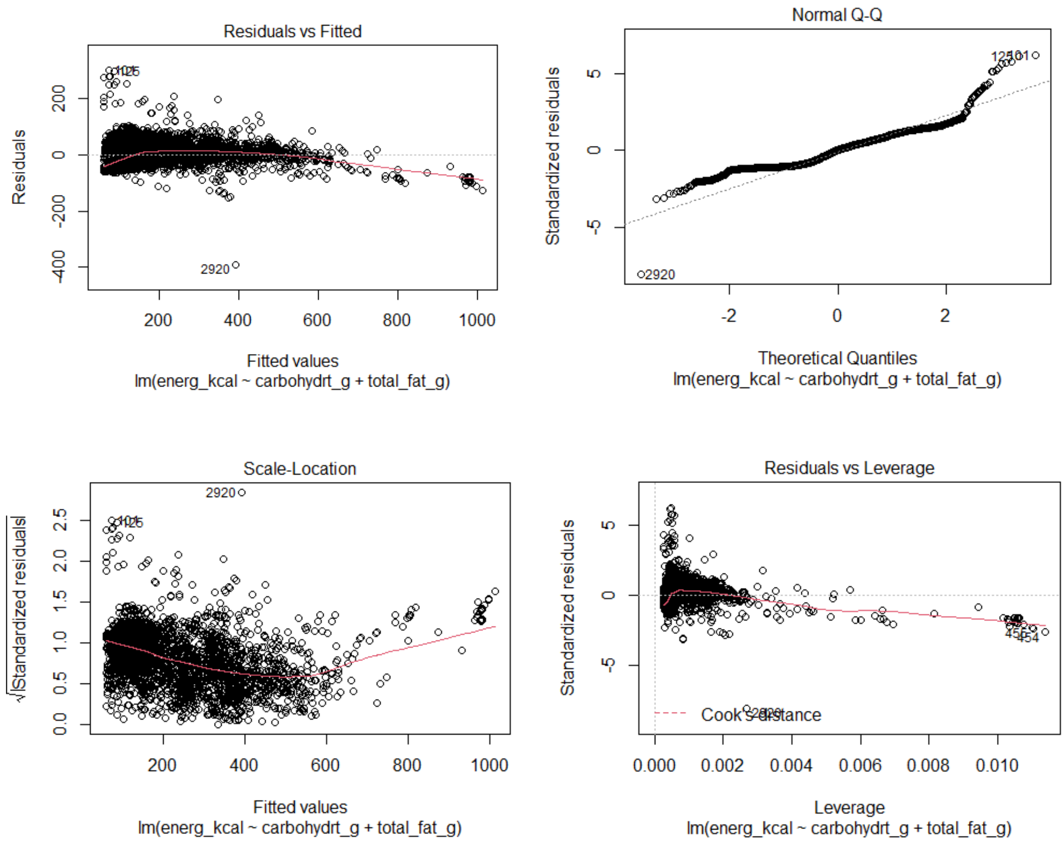
Independence: Similar as above.

Normality: Q-Q plot suggests problems with normality, the tails deviate a lot from the line.

Linearity: The Residuals vs Fitted plot shows that the points are generally randomly placed, and the red line seems horizontal. This suggests linearity.

Constant variance: The Scale-Location plot does not have a horizontal red line, and this might suggest an increase in variance at some point around the centre of the data.

Returning to the initial question:



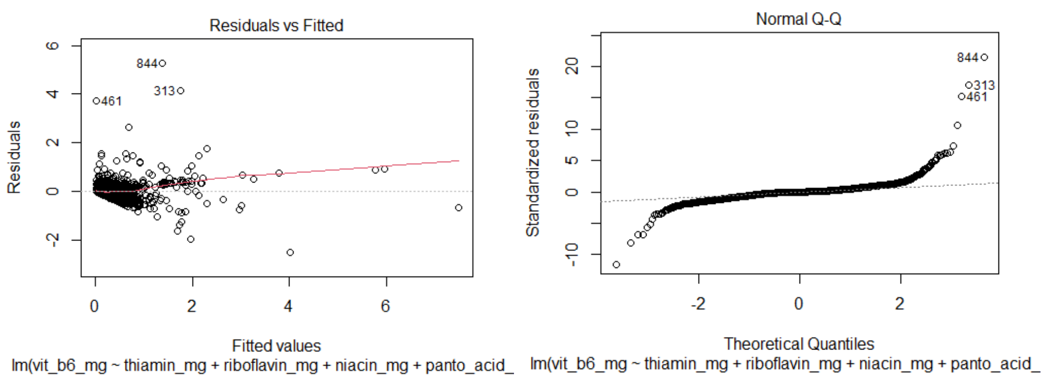
Independence: Similar to what was said for Question 1.

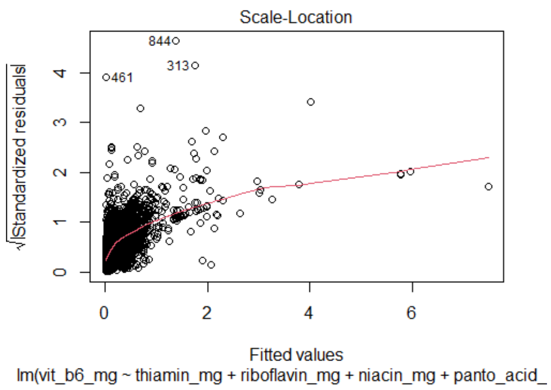
Normality: Again, there are some problems with normality as we can see the tails deviating. Generally the data in question is not normally distributed.

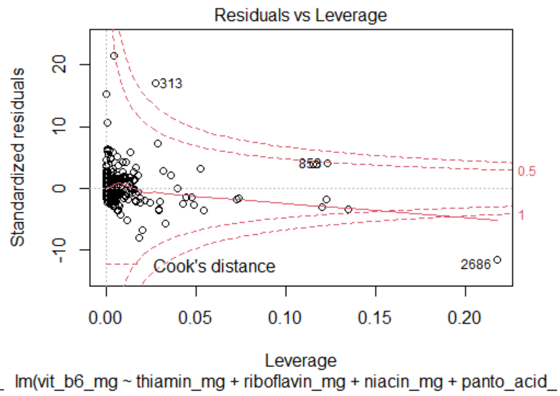
Linearity: In the Residuals vs Fitted plot the points are randomly placed around the red line, and the red line is generally horizontal so linearity can be a valid assumption.

Constant variance: In the Scale-Location plot, the red line is curved in the middle, indicating that variance is not constant and changes as values increase.

B vitamins regression:







Independence: For the same reasons as before, independency fails.

Normality: The Q-Q plot suggests that the data is not normally distributed, since there is a lot of deviation from the line.

Linearity: The red line in the Residuals vs Fitted plot has a positive slope, which indicates that the relationship of the variables is not linear, and another model could be a better fit.

Constant variance: The variance is not constant across the data, and we can see that from the red lines in the Residuals vs Fitted and Scale-Location plots. The red lines are not horizontal and it seems like variance increases when values increase.

There is a number of possible violations of linear regression model assumptions in the way that I have fitted the data. Violations in the assumptions of normality and constant variance can interfere with the generation of errors, and consequently cause wrong results for the confidence and prediction intervals. That would be problematic if we use these linear models to make predictions for other data points. The predictions rely on the existence of a normal distribution with constant variance.

On the other hand, violations that occur in linearity and independence can seriously affect the coefficients and the intercept of the model. If the data is not related linearly, or if there are many outliers, then the coefficients will not be correct. The data can be related in an exponential or quadratic manner for example, which would skew the coefficients much. In addition, absence of independece means that there is a common factor that can significantly influence changes in the dependent variable, which is not necessarily related to the independent variables, or affects some outcomes more than other.

Because of all these violations, if I had more time to work on the project I would have tried logarithmic transformations on my variables, and this might have improved normality, variance and linearity. Especially for the question regarding how B vitamins are related, I would have tried various different models that would possibly explain it better, such as polynomial regression. ANOVA would help in order to choose the more suitable model.

There were also some questions that came to mind that I did not have the time to investigate. While going through scientific bibliography, I learned that fat-soluble vitamins, A, K, D and E are being found and stored in lipids [1][2]. I would have liked to examine the relationship between the lipid content of foods and the amount of fat soluble vitamins they contain. More generally, the dataset has many different nutritional variables and there is a lot of potential of discovering interesting relationships that I did not manage to see.

Conclusion

To summarise, the analysis that was done tried to highlight the potential influence of fats, carbohydrates and sugar on the number of calories a food contains. It was shown that it is very probable that a linear relationship between these substances and energy exists, and it is positive, although there were violations that may make the results inaccurate. Furthermore, there was an investigation of the relationship between the vitamins in the B complex. A very insignificant linear relationship was found, with weak results and many possible assumption violations. The model does not sufficiently overcome randomness or other complex relationships that can occur in the science of nutrition, and better models could be fitted to describe the reality of these nutrients.

References

Primary Source – Dataset

Awram, C. (2016). Food Nutritional Values. [online] Available at: https://data.world/awram/food-nutritional-values [Accessed 6 Nov. 2024].

Secondary sources

[1]. Combs, G.F. (2008). *The Vitamins: Fundamental Aspects in Nutrition and Health*. Third ed. London: Elsevier Academic Press, p.64.

[2]. Gropper, S.S., Smith, J.L. and Groff, J.L. (2009). *Advanced Nutrition and Human Metabolism*. Fifth ed. Belmont: Wadsworth, Cengage Learning, pp.323–369, 373.

[3]. National Institutes of Health. (2024). *Vitamin B12*. [online] Available at: https://ods.od.nih.gov/factsheets/VitaminB12-HealthProfessional/ [Accessed 6 Nov. 2024].

[4]. US Department of Agriculture, Agricultural Research Service and Nutrient Data Laboratory (2016). *Composition of Foods Raw, Processed, Prepared Release 28*. [online] *USDA National Nutrient Database for Standard Reference*, US Department of Agriculture, pp.1-26. Available at: http://www.ars.usda.gov/ba/bhnrc/ndl [Accessed 6 Nov. 2024].