INVESTIGATING THE NUTRIENT COMPONENTS THAT ARE MOST IMPACTED BY LOW-CALORIE DIETS

Thuy Linh (Kylie) Le

COMP20008 - Assignment 2

May 19, 2023

Table of Contents

1/ Aim

Pepsi's diet cola, Nestle's fat-free cocoa, or Kellogg's whole-grain breakfast cereal are only a few examples of companies introducing reduced-calorie products in response to consumers' demand. Reducing calories, a measure of energy in food and beverages, has gained considerable attention and can be a step towards a healthier lifestyle. However, low-calorie diets may not address overall nutritional needs and should be exercised with advice from professionals.

This project will investigate potential relationships between the amount of calories and the amount of other components in food. It aims to identify nutrient components that are most inadvertently affected by the reduction of calorie intake. This may give insights into the appropriateness of low-calorie diets for people with certain health conditions.

This project targets people who are looking to lose weight or maintain a healthy weight, where reducing calorie intake is often a key strategy. Target audiences also include community health workers who wish to promote balanced diets and food manufacturers who wish to tailor their low-calorie products to better address other nutritional needs of customers.

2/ Background

2.1/ Dataset

The dataset utilized in this project is the nutrient file from Australian Food Composition Database - Release 2 (Foodstandards.gov.au, 2016). It includes information of 1616 foods and beverages per 100 grams portion. Each food has a unique public food key as a string of letters and numbers, followed by a 5-digit numeric series representing the classification. This is followed by a string of Food Name, and, lastly, numeric values of up to 290 nutrients and food components contained.

2.2/ Target variable and Explanatory variables

'Energy with dietary fibre, equated (kJ)' will be referred to as the target variable throughout this report. It is calculated as follows:

Total energy = 17(Total protein) + 37(Total fat) + 17(Carbohydrate) + 37(Total dietary fibre)

(where total energy is measured in kilojoules per serve and other variables are measured in grams per serve) (www.mydailyintake.net, n.d.)

From this definition, the relationships between energy and protein, fat, carbohydrate, and fibre are already clear. Hence, variables relating to these nutrients will not be examined in this project. 283 other nutrients and food components are used as potential explanatory variables.

3/ Data Pre-processing

The dataset has some drawbacks: (1) a large number of missing values, and (2) different features' measurements and ranges. Before addressing these problems, regular expression is used to clean newline characters contained in features' names.

3.1/ Feature removal

As discussed in part 2.2, these features are removed: 'Energy, without dietary fibre, equated (kJ)', 'Protein (g)', 'Fat, total (g)', 'Total dietary fibre (g)', 'Available carbohydrate, without sugar alcohols (g)', 'Available carbohydrate, with sugar alcohols (g)'. 'Public Food Key', 'Food Name', and 'Classification' are also removed as this project's focus is on the relations between nutrient components.

All features having more than two-thirds of entries missing are removed as imputation techniques require a certain amount of data presented in order to work reasonably well. It is notable that lists of features that break the above threshold are constructed separately for train and test data. The lists are then merged, and features in the merged list are removed for both sets. This avoids the effect of "unlucky" splits where one set contains the majority of missing data, causing inaccurate separate imputation.

3.2/ Normalization and Imputation

Since the features have different units and a significant amount of extreme values and large ranges, MinMaxScaler is used to rescale the values to be between 0 and 1.

Foods with similar amounts of other nutrients are likely to contain similar amounts of the nutrient with missing values. Hence, the chosen technique is K-nearest neighbors (KNN) imputation.

4/ Feature Selection

The pre-processed dataset still contains more than 70 features. Redundancy induces extra computational effort or overfitting, so reducing the dimensionality is needed.

4.1/ Zero-variance removal and Power transform

Variables with zero variance are removed since no relationship between these variables and the target can be deduced.

Pearson correlation assumes that variables are roughly normally distributed. Since the data is heavily skewed, Power Transformer needs to be applied feature-wise to make data more Gaussian-like.

4.2/ Correlation

Since all features are continuous, Pearson correlation is chosen to assess their collinearity. Predictors that are highly correlated with other predictors or slightly correlated with the target are deemed insignificant for the models and removed. Correlation matrices and scatter graph matrices are generated to visualize the collinearity of selected features (figure 1, 2).

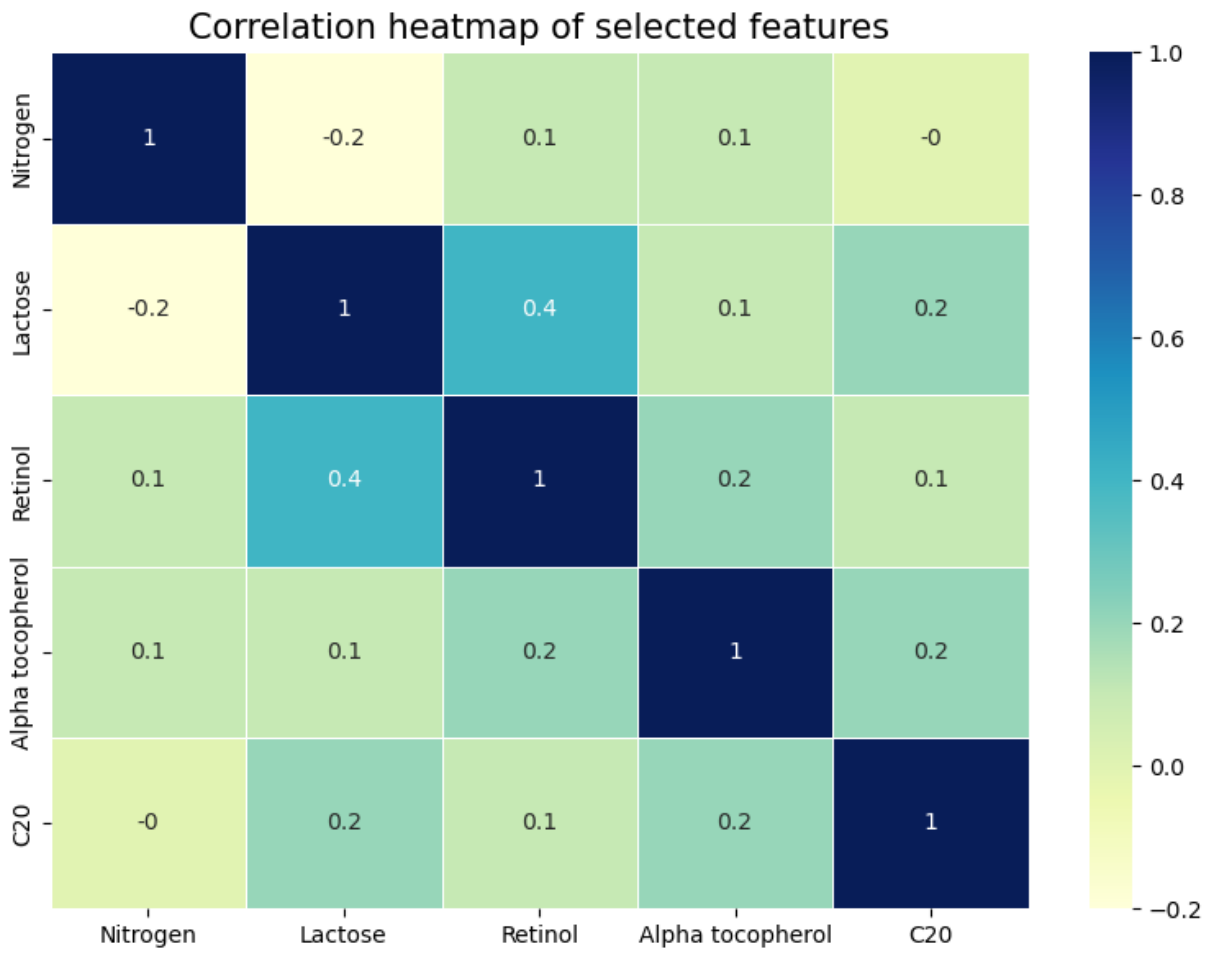

Figure 1: Correlation heatmap of selected features

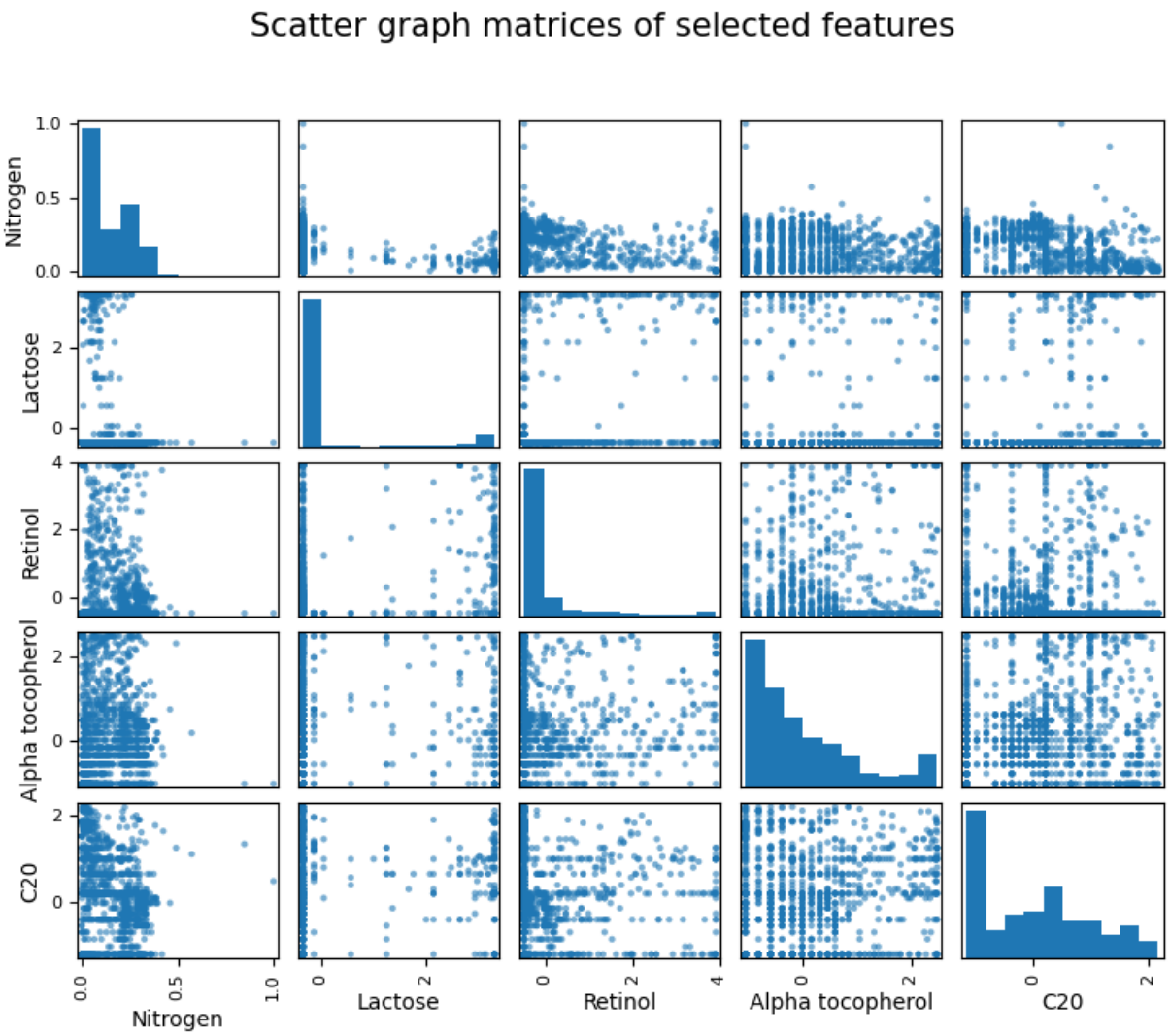

Figure 2: Scatter graph matrices of selected features

5/ Regression model

5.2/ Linear model

This model examines the relationship between 5 predictors ('Nitrogen', 'Lactose', 'Retinol', 'Alpha tocopherol', 'C20') and the response variable.

10-fold Cross Validation (CV) is used to:

(1) give the average interception and coefficients of the linear regression model, allowing for more reliable interpretation of the relation between predictors and the target variable

(2) evaluate the choice of above predictors using 3 metrics: R-squared on train set, R-squared on test set, and Mean Squared Error (MSE) on test set, noting that MSE is calculated on scaled data (table 1).

The obtained linear regression expression is as follows:

$$Energy = 0.15 + 0.26(Nitrogen) + 0.55(Lactose) + 0.12\ (Retinol)$$

$$+ 1.27(Alpha\ tocopherol) + 0.15(C20)$$

5.4/ Regression tree

The chosen features contain outliers (figure 4), to which linear regression is sensitive. Regression tree should be utilized as it is not largely influenced by outliers. 10-fold CV on a train set is used to determine the tree depth minimizing MSE. Figure 3 implies that 3 may be an optimal depth where MSE is small and complexity level is acceptable. A regression tree is then visualized (figure 4). 10-fold CV is performed again to evaluate the performance of the tree regression model with max depth 3 using MSE metric (table 1).

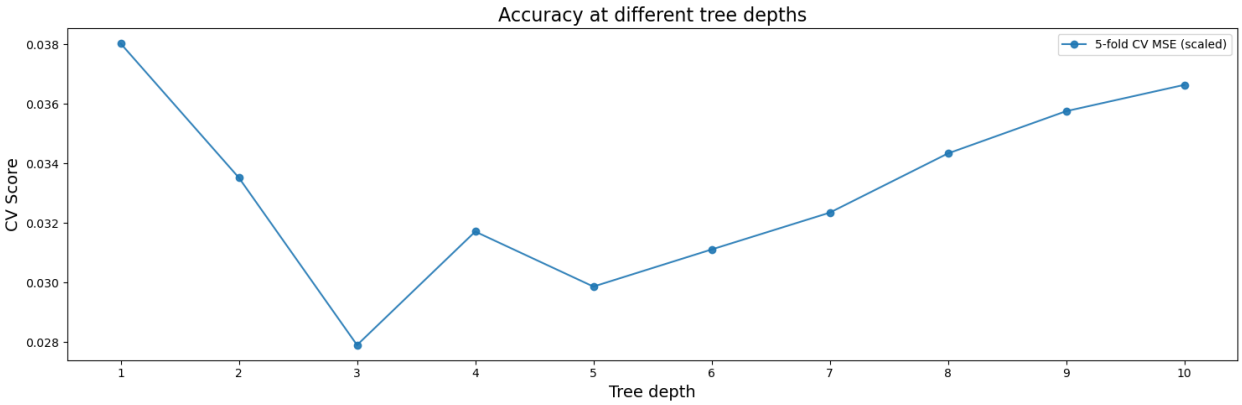

Figure 3: Average MSE values for different tree depths

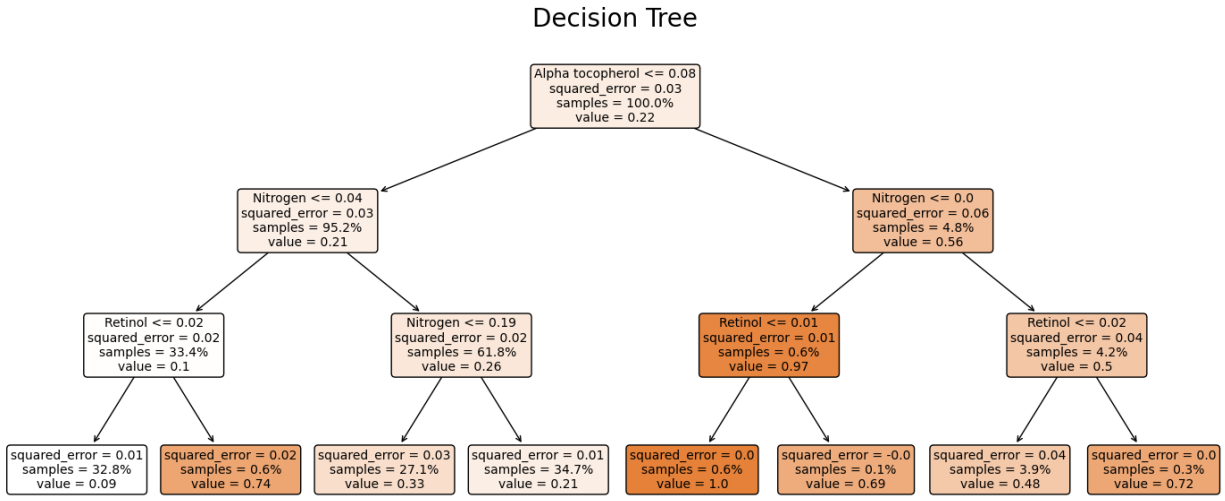

Figure 4: Regression tree

6/ Discussion

6.1/ Evaluation

	10-fold CV Train R-squared	10-fold CV Test R-squared	10-fold CV MSE (scaled)
Linear model	0.23	-0.13	0.04
Regression Tree	n/a	n/a	0.03

Table 1: Goodness of fit

The residuals for the Linear model are more not randomly distributed (figure 5), so a linear relationship may not exist between the predictors and the target. The average R-squared on train data of the Linear model is low (0.23). In addition, the average R-squared on test sets is negative (-0.13). Hence, the Linear model does not generalize well, performing better on training data but worse than the mean of the target on test data. This overfitting problem may be due to noisy data.

The Regression tree may be a more suitable model since it has a lower average MSE (0.03) than the Linear model (0.04). This is potentially due to the fact that tree regression is not as sensitive to outliers and over simplicity as linear regression.

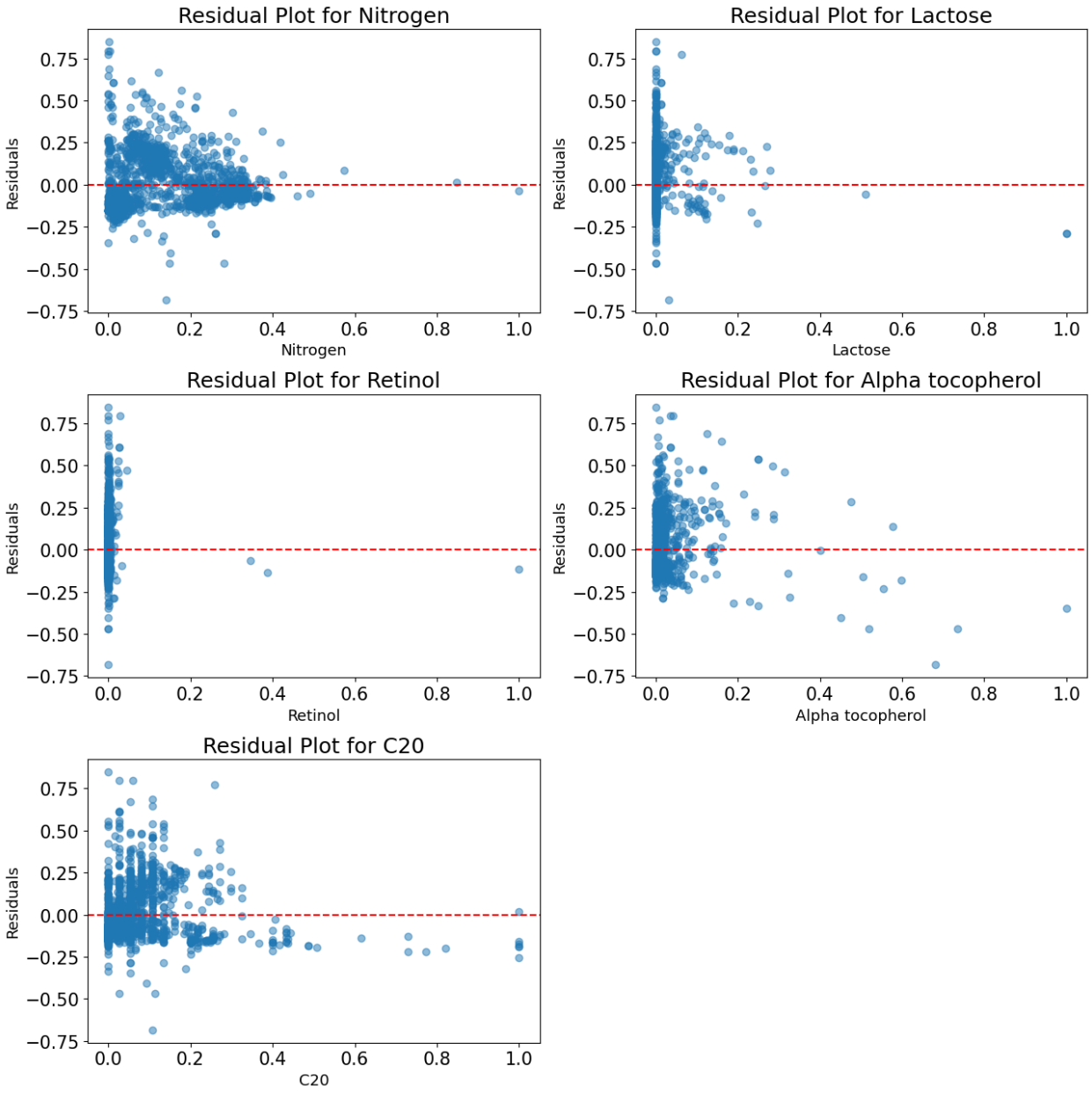

Figure 5: Residual plots for Linear model

6.2/ Conclusion

The key variables for splitting in the Regression tree are 'Alpha tocopherol', 'Nitrogen', 'Retinol' (figure 4). The coefficients of these key variables in the Linear model reveal positive relationships. Hence, among all listed nutrients, the decrease in energy most likely coincides with the decrease in vitamin E (alpha tocopherol), vitamin A (retinol), and nitrogen.

Special attention should be paid to beneficial nutrients while constructing low-calorie diets. For example, people with inflammatory conditions, risks of heart or eye disease, or skin issues may avoid low-calorie diets as vitamin E and vitamin A intakes become important. From these findings, food manufacturers can focus on nutritional supplement products targeting customers who are following low-calorie meal plans. Community health workers can raise special attention to the mentioned nutrients, which can make low-calorie diets more balanced.

6.3/ Limitation and Improvement

KNN Imputation assumes that foods with similar amounts of other nutrients will have similar amounts of the nutrient we are trying to impute. Imputation can give more accurate results if the assumption is limited to foods of the same group. However, one limitation of the dataset is that some food groups only have a small number of foods presented (for example, special dietary foods (ID 12) with 2 items). Within such groups, it is much more likely for a feature to contain all NaN values, making imputation infeasible.

Many features have extreme values and are heavily right-skewed. After partly addressing this problem by Power Transform, plots on the diagonal of figure 2 still show skewed distribution, which hinders the performance of linear regression models. To improve, Interquartile Range (IQR) can be used to determine the boundaries, then removal or imputation can be performed to handle outliers.

The average R-squared on test sets of the Linear model is negative and the Regression Tree has the better performance compared to the Linear model. This suggests that non-linear relationships between explanatory variables and the target. Non-linear regression should be used to give a more accurate model.

Reference list

Foodstandards.gov.au. (2016). *Download Excel files (Australian Food Composition Database - Release 1)*. [online] Available at: https://www.foodstandards.gov.au/science/monitoringnutrients/afcd/Pages/downloadableexcelfile s.aspx.

www.foodstandards.gov.au. (n.d.). *Classification of foods and dietary supplements*. [online] Available at: https://www.foodstandards.gov.au/science/monitoringnutrients/ausnut/classificationofsupps/Pages/default.aspx.

www.mydailyintake.net. (n.d.). *Daily Intake Guide: Healthy eating, made easy. Front-of-pack labelling for food and drink in Australia. - Calculating Food Energy*. [online] Available at: http://www.mydailyintake.net/calculating-energy/.