ST309 Group Project Report

Assessing the factors that lead to obesity

16923

17402

17915

(All participants contributed equally)

Table of contents

1. Introduction

1.1 Description of the problem

"At least 2.8 million people die each year as a result of being overweight or obese".

-World Health Organisation (WHO)-

The WHO defines overweight and obesity as abnormal or excessive fat accumulation which leads to increased health risks. As per the European Association for the Study of Obesity (EASO), obesity has become a global epidemic which has nearly doubled since 1980. Moreover, overweight and obesity are the fifth leading cause for worldwide deaths. However, obesity is preventable! Therefore, the objective of this project is to identify and analyse the factors that contributed towards obesity of individuals and use this information to provide tailored advice to combat the same.

Overweight and obesity are classified according to an individual's BMI. An individual's BMI can be calculated as follows:

$$\text{Body mass Index (BMI)} = \frac{weight\ (kg)}{height^2 (m^2)}$$

According to the World Health Organisation (WHO), an individual's BMI can be divided into 4 categories:

- Underweight - BMI less than 18.5 ($BMI < 18.5$)
- Normal - BMI in the range of 18.5 to 24.9 ($18.5 \leq BMI < 24.9$)
- Overweight - BMI in the range of 25.0 to 29.9 ($25 \leq BMI < 30$)
- Obesity - BMI equal to or greater than 30.0 ($BMI \geq 30$)

1.2 Using data analytics

Identifying, analysing and providing feedback regarding the factors that affect obesity can be classified as a data analytic problem for two reasons. The first reason is that a data set can be analysed in order to draw inferences from it. Secondly, statistical procedures such as multiple linear regressions, decision trees, logistic regressions etc. can be used to manipulate the dataset to explore trends, correlation and causality among the variables.

This report is an example of diagnostic, predictive and prescriptive analytics because it analyses the factors that cause obesity, uses several variables to predict obesity and provides tailored advice suggesting how to combat the same.

1.3 Previous studies related to the topic

Multiple studies have been conducted to identify the causes of obesity as regularly updated research is vital in capturing the factors leading to obesity.

Hruby et al. (2016) stated that dietary and lifestyle factors may lead to obesity. Beverages containing high sugar content, a poor diet, physical inactivity, extended screen time and a lack of sleep were recognised as the most prominent dietary and lifestyle factors that lead to obesity. However, even if a group of individuals have similar dietary and lifestyle choices, they may exhibit a varying propensity to obesity because of biological factors such as age and sex (Dhurandhar et al., 2014; Valera et al., 2015)

Further, studies such as Elks et al. (2012) discovered that the genetic composition of an individual may lead to obesity. On a scale of 0 to 1 (where 0 indicates that most of the variation in a particular characteristic between different individuals is not genetic and 1 indicates that genetics explains all the variation), the heritability estimate of obesity among twins and families in the sample were between 0.25 and 0.9. This suggests that obesity may be genetic. Moreover, other studies have concluded that the genetic risk associated with obesity increased due to lifestyle and dietary choices such as increased intake of sugar sweetened beverages (Qi et al., 2012) and regular fried food consumption (Qi et al., 2014).

2. Data

2.1 Data description and data source

The dataset used for the analysis was the estimation of obesity levels based on a collection of information on eating habits and physical condition data obtained from the UCI machine learning repository (Refer 6.1 for further information). This dataset contains information regarding 17 attributes of 2111 individuals between the ages of 14 and 61 from Mexico, Peru and Columbia. Some variables that were not relevant for the analysis were removed and this will be explained further in 2.2 Data Preparation. 23% of the data was collected from respondents via an internet platform and 77% of the data was generated using the SMOTE filter in the WEKA tool.

Table 1: Summary of the relevant variables in the dataset and the associated levels

Category	**Variable and description**	**Variable Type**
Dependent Variable	BMI	Numeric
Respondent Characteristic	Gender	Factor with 2 levels: Female and Male
Respondent Characteristic	Age (in years)	Numeric
Respondent Characteristic	Family.history.with.overweight	Factor with 2 levels: Yes and No
Eating habits	Fast.food.intake	Factor with 2 levels: Yes and No

Eating habits	Vegetable.consumption.freq	Factor with 3 levels: Never, Sometimes and Always
Eating habits	No.main.meals	Factor with 4 levels: One, Two, Three and Four
Eating habits	Food.between.meals	Factor with 4 levels: No, Sometimes, Frequently and Always
Eating habits	Daily.liquid.intake (in litres)	Factor with 3 levels: Less than a litre, 1-2 litres and More than 2 litres
Eating habits	Calc.daily.calories	Factor with 2 levels: Yes and No
Physical condition	Physical.activity.freq	Factor with 4 levels: Never, 1-2 days, 2-4 days, and 4-5 days
Physical condition	Smoke	Factor with 2 levels: Yes and No
Physical condition	Alc.consumption	Factor with 4 levels: No, Sometimes, Frequently and Always
Physical condition	Transportation.used	Factor with 5 levels: Automobile, Bike, Motorbike, Public Transport and Walking
Physical condition	Time.on.tech	Factor with 3 levels: 0-2 hours, 3-5 hours, and more than 5 hours

Approximately 73% of the dataset consisted of individuals who were either overweight or obese (Refer Table 2).

Table 2: Summary of the proportions of underweight, normal, overweight and obese individuals in the dataset (sample size = 2111)

BMI	**BMI < 18.5 (Underweight)**	**18.5 ≤ BMI < 25 (Normal)**	**25 ≤ BMI < 30 (Overweight)**	**BMI ≥ 30 (Obese)**
Number of people in data within this range	271 (13% of the total data)	297 (14% of the total data)	570 (27% of the total data)	974 (46% of the total data)

2.2 Data Preparation

Before conducting an exploratory data analysis, the data was tidied. Variables such as Weight and Height were removed from the dataset and used to create the variable BMI. The variable Weight class was also removed as it was based on BMI. The nominal variables wrongly coded as numbers were converted back to categorical variables (Refer 2.2.1). The

levels of some predictors were merged (Refer 2.2.2) and the base level of required predictors were changed (Refer 2.2.3). Further, the summary of data revealed that there were no missing values in the dataset.

2.2.1 Converting predictors incorrectly coded as numerical into factors

As per the questionnaire used to collect the data (Refer 6.2), the following predictors are categorical variables. While the data collected via the online platform was coded as integers, most of the data generated using the SMOTE filter in the Weka tool was coded as decimal numbers. Therefore, these predictors were incorrectly identified as continuous variables. These predictors were then grouped into levels according to the initial questionnaire used to collect information (Refer 6.2).

- Vegetable consumption frequency (Vegetable.consumption.freq)
  This variable was grouped into levels as follows:
  Never - Vegetable.consumption.freq < 1.5
  Sometimes - Vegetable.consumption.freq ≥ 1.5 & Vegetable.consumption.freq < 2.5
  Always - Vegetable.consumption.freq ≥ 2.5

- Number of main meals (No.main.meals)
  The number of main meals should be a positive integer. Therefore, the values were grouped accordingly:
  One - No.main.meals < 1.5
  Two - No.main.meals ≥ 1.5 & No.main.meals < 2.5
  Three - No.main.meals ≥ 2.5 & No.main.meals < 3.5
  Four - No.main.meals ≥ 3.5

- Physical activity frequency (Physical.activity.freq)
  As per the questionnaire this variable was grouped as below:
  Never - Physical.activity.freq < 0.5
  1 - 2 days - Physical.activity.freq ≥ 1.5 & Physical.activity.freq < 2.5
  2 - 4 days - Physical.activity.freq ≥ 0.5 & Physical.activity.freq < 1.5
  4 - 5 days - Physical.activity.freq ≥ 2.5 & Physical.activity.freq < 3.5

- Time on tech (Time.on.tech)
  This variable was separated into levels as mentioned below:
  0 - 2 hours - Time.on.tech < 0.5
  3 - 5 hours - Time.on.tech ≥ 0.5 & Time.on.tech < 1.5
  More than 5 hours - Time.on.tech ≥ 1.5

- Daily liquid intake (Daily.liquid.intake)
  This was grouped as follows:
  Less than a litre - Daily.liquid.intake < 1.5
  Between 1 and 2 litres - Daily.liquid.intake ≥ 1.5 & Daily.liquid.intake < 2.5

More than 2 litres - Daily.liquid.intake ≥ 2.5

2.2.2 Merging levels of categorical predictors

The levels of the following categorical predictors were merged based on the boxplot of the predictor against BMI and the frequency of the level. After observing the boxplot of the predictor vs BMI, levels with similar medians and quartiles were identified. If these identified levels had a low frequency (a few number of observations) they were considered for merging.

- Alcohol Consumption (Alc.consumption)

Originally the factor Alcohol consumption had 4 levels: 'Always', 'Frequently', 'Sometimes' and 'Never'. However, only one individual had chosen the option 'Always'. Therefore, this level was merged with 'Frequently'. After merging levels, Alcohol consumption had three distinct levels: 'Frequently', 'Sometimes' and 'Never'.

- Transportation used (Transportation.used)

Originally this factor had 5 different levels. The similarity of the methods of transport were also considered when merging levels. 'Walking' and 'Bike' had very similar medians. They are also forms of cardio. Therefore, these levels were merged to form the new level 'Walk/Bike'. Similarly, 'Automobile' and 'Motorbike' were merged to form the level 'Personal vehicle'. After merging levels this factor had 3 levels: 'Walk/Bike', 'Personal vehicle' and 'Public transport'.

2.2.3 Changing the base level of predictors

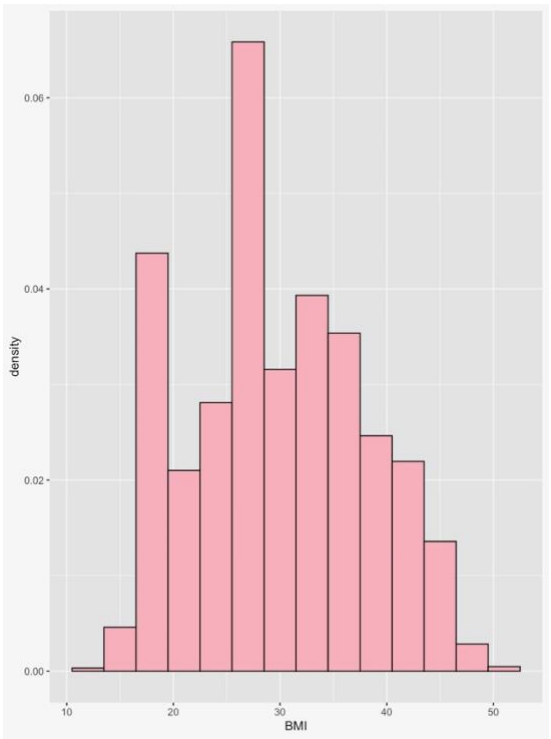
The base level of a predictor is defined as the level of comparison for all levels of the predictor. The default base level of a categorical predictor is the level that occurs first alphabetically. The new base levels for all the categorical predictors were set to their modal levels (most frequently occurring levels). Practicality of the new base level was also considered. Therefore, the base levels of the following predictors were changed.

- Number of main meals was changed to 3
- Vegetable consumption frequency was changed to Sometimes
- Gender was changed to Male
- Family history with overweight was changed to Yes
- Fast food intake was changed to Yes
- Food between meals was changed to Sometimes
- Alcohol consumption was changed to Sometimes
- Transportation used was changed to Public transport

2.3 Variables and relationships among variables

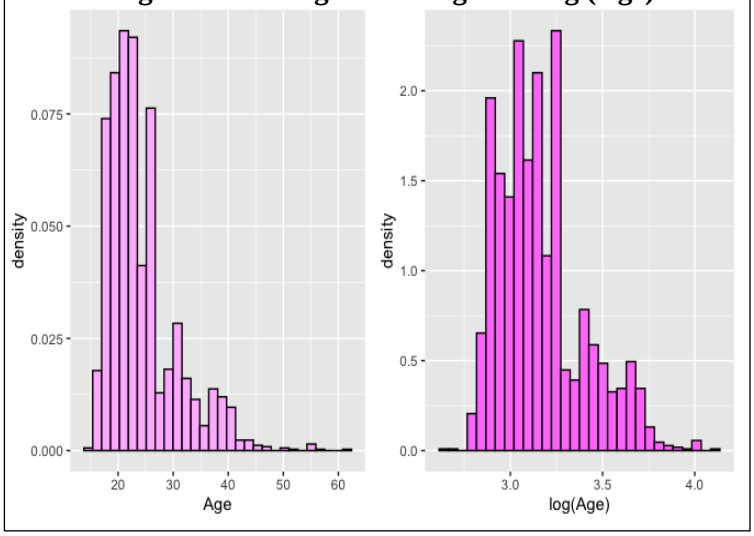
Firstly, a histogram was plotted to observe the nature of the dependent variable (BMI). Figure 1 reveals that the values for BMI are normally distributed with a mean value (29.70) approximately equal to the median (28.72). Further, the values for BMI ranged from 13.00 to 50.81. There are also additional peaks observed in Figure 1.

Figure 1 – Histogram of BMI



Next, histograms were plotted to observe continuous variables and bar charts were used to plot categorical predictors. A log transform (with respect to base e) was applied to Age and this transformation was used in the multiple linear regression (MLR) for 2 reasons. Firstly, when conducting a univariate analysis of Age, the data was moderately positively skewed as the mean value of Age (24.31) was less than its median (22.78). After a log transform, the data had less of a positive skew (Refer Figure 2). Secondly, the minimum value of age was 14 and this is greater than 0. This transformation was not necessary for the logistic regression and trees as they do not require predictors to follow a normal distribution.

Figure 2 – Histograms of Age and log (Age)

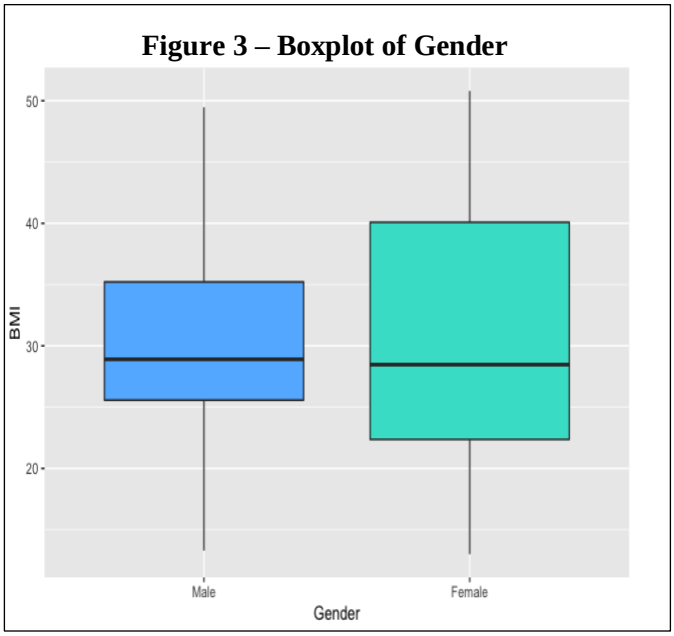


It was difficult to observe potential relationships among variables by using a scatterplot matrix as most variables were categorical. Similarly, a correlation matrix could not be used to discover the correlation among variables.

The relationship between the predictor variables and BMI were explored next. This was done by creating scatterplots for the continuous predictors vs BMI and plotting boxplots for the categorical predictors against BMI.

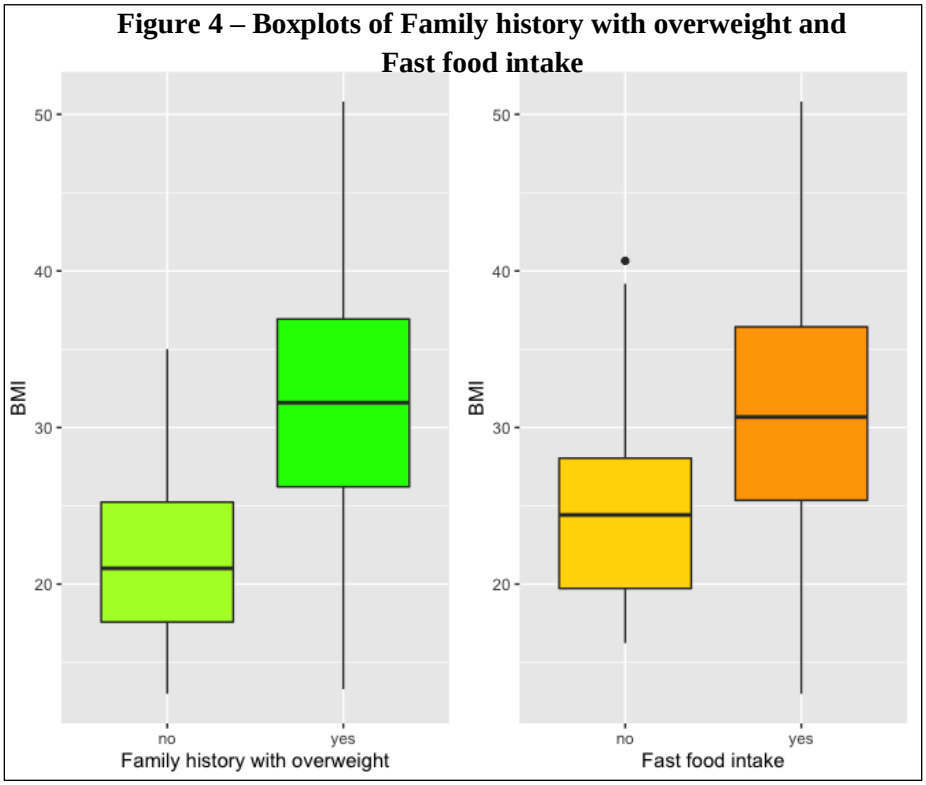
From Figure 3, gender on its own had little predictive power for BMI. However, when combined with other predictors it may be helpful to predict BMI.

Figure 3 – Boxplot of Gender



As per Figure 4, factors such as Family history with overweight and Fast-food intake positively correlated with BMI. According to Taylor (2011), research studies estimate that obesity is about 40% hereditary (genetic) and 60% due to environmental factors. However, there is a large amount of variation in scientific literature. Further Mandal (2019) mentions that an increased intake in fast food is associated with an increased risk of obesity, weight gain and less successful weight-loss maintenance. Further, as expected the level of physical fitness of an individual negatively correlated with BMI.

Figure 4 – Boxplots of Family history with overweight and Fast food intake



3. Data analysis using statistical learning procedures

3.1 Multiple Linear Regression (MLR) to predict BMI

3.1.1 MLR to predict BMI of individuals greater than 18.5 (Model 1)

The Multiple Linear Regression (MLR) explores which factors contribute to obesity (BMI exceeding 30) and the magnitude of the impact of these factors on obesity. Individuals who were underweight (BMI less than 18.5) were excluded from the model as they may have a different distribution and therefore distort the model.

The MLR was formed using the best subsets regression method. A training set consisting of 1056 datapoints was used to create the model and the remaining data (1055 data points) was used as a testing set to assess the predictive power of the model. The best subsets method calculates the AIC (Akaike information criterion) with every possible combination of the predictors and identifies the model that minimizes AIC. The first model created using best subsets consisted of all predictors bar Smoke, Gender and Physical activity frequency. The residual assumptions were not violated. All data points had a Cook's distance equal to or less than 0.02 except for point 218 which had a Cook's distance of 0.04.

Although the Cook's distance of this point was low, it was significantly higher than the other points. Therefore, it was investigated. However, re-running the model without this point (218) did not change the model. This implied that datapoint 218 was not a leverage point and thus, should not be removed.

Figure 5 – Regression table of Model 1

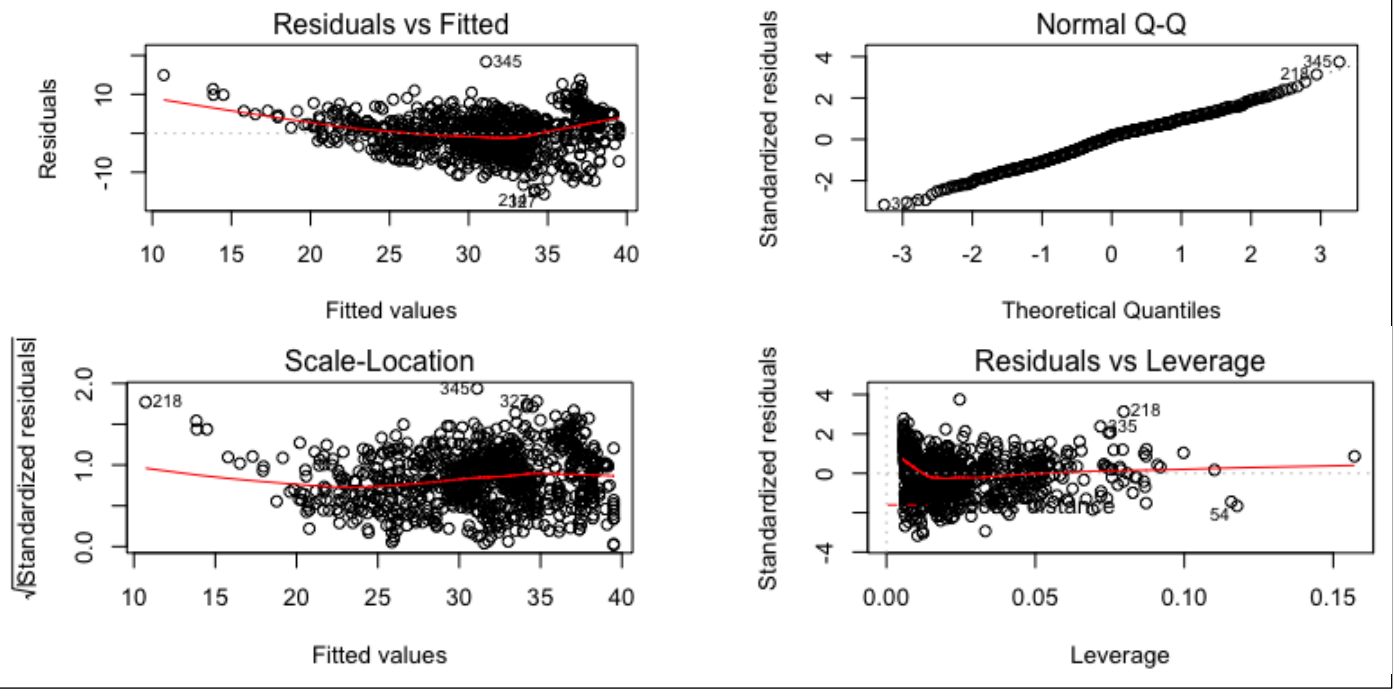
	BMI		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	13.79	7.61 – 19.98	**<0.001**
Family.history.with.overweight [no]	-5.51	-6.59 – -4.42	**<0.001**
Fast.food.intake [no]	-3.31	-4.42 – -2.21	**<0.001**
Vegetable.consumption.freq [Never]	1.74	0.01 – 3.46	**0.049**
Vegetable.consumption.freq [Always]	4.10	3.41 – 4.79	**<0.001**
No.main.meals [One]	-2.57	-3.56 – -1.58	**<0.001**
No.main.meals [Two]	-1.37	-2.56 – -0.18	**0.024**
No.main.meals [Four]	-0.88	-2.58 – 0.81	0.307
Food.between.meals [Always]	-3.48	-5.41 – -1.55	**<0.001**
Food.between.meals [Frequently]	-5.02	-6.37 – -3.67	**<0.001**
Food.between.meals [no]	-1.14	-3.46 – 1.18	0.336
Daily.liquid.intake [Less than a litre]	-0.77	-1.62 – 0.08	0.077
Daily.liquid.intake [More than 2 litres]	1.23	0.43 – 2.03	**0.003**
Calc.daily.calories [yes]	-3.16	-4.94 – -1.37	**0.001**
Time.on.tech3-5 hours	1.35	0.63 – 2.07	**<0.001**
Time.on.tech [More than 5 hours]	0.14	-1.03 – 1.31	0.815
Alc.consumption [Frequently]	-3.29	-4.98 – -1.60	**<0.001**
Alc.consumption [Never]	-1.51	-2.29 – -0.72	**<0.001**
Transportation.used [Personal Vehicle]	-3.14	-4.16 – -2.13	**<0.001**
Transportation.used [Walk/Bike]	-4.30	-6.14 – -2.45	**<0.001**
log_Age	5.84	3.90 – 7.78	**<0.001**
Observations	917		
R2 / R2 adjusted	0.502 / 0.491		

Model 1 has an intercept of 13.79. This means that an individual has an expected BMI of 13.79 when all predictors equal their base level (Refer 2.2.3 for the base levels). Further, the baseline for log (Age) is zero. The coefficient estimate of all predictors except log (Age) is the amount by which the expected BMI will increase/decrease when the predictor deviates from the base level. For example, the base level of the predictor Daily liquid intake was 1-2 litres. Therefore, there is a 0.77 expected decrease in BMI if an individual consumed less than a litre of daily liquids and a 1.23 expected increase in BMI if an individual consumed more than 2 litres of daily liquid compared to if they consumed 1-2 litres of daily liquids. The coefficient estimate corresponding to log (age) is 5.84 and this means that a 1% increase in age will lead to a 0.0584 increase in expected BMI.

The Residuals vs Fitted plot (Figure 6 – Top left) was slightly funnel shaped, but this wasn't an issue. Normality of the residuals was observed as the Normal Q-Q plot (Figure 6 – Top right) lies along the x-y diagonal. The Scale-Location plot (Figure 6 – Bottom left) is pattern less. This reveals that there are homogeneous variances among different observations and that the fitting is adequate. The Residuals vs Leverage plot (Figure 6 – Bottom right) reveals that there were no unusually large outliers (bad leverage points). Further, all points had a low Cook's distance.

There was no multicollinearity between the predictors in the model because the vif (variance inflation factor) for all predictors was between 1 and 2.

Figure 6 – Diagnostic plots of Model 1



The testing set was then used to assess the predictive power of the model. As observed in Figure 7, the predicted values of BMI were plotted against the actual value of BMI. If the model perfectly predicted the data, all datapoints would lie along the line y = x.

Figure 7 – Actual vs Predicted BMI (Model 1)

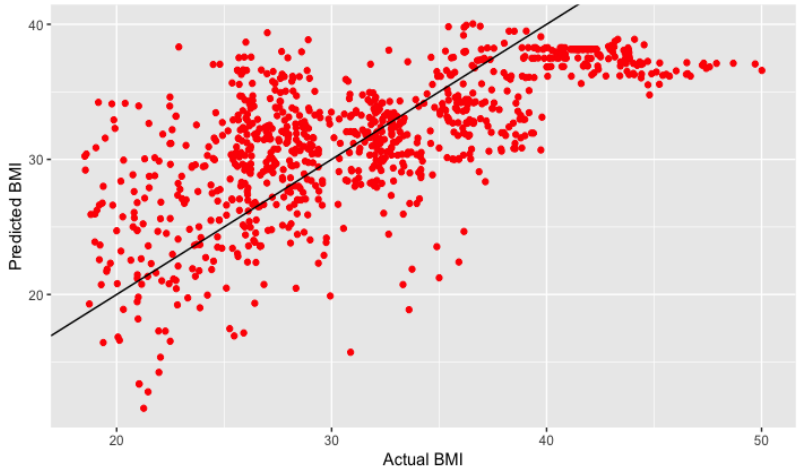


Figure 7 indicates that the Model 1 predicted BMI up to 40 well. However, this model cannot be used to predict BMI above 40. Accordingly, the data was split into two; individuals with a BMI greater than 40 and those under 40 (no individual had a BMI equalling 40). Using these subsets, 2 further regression models were run.

3.1.2 MLR to predict BMI less than 40 (Model 2 - Part A)

When a multiple linear regression was conducted using the best subsets method on the training set (containing of 785 data points), the model contained 8 predictors. Namely, Gender, Family history with overweight, Fast food intake, Vegetable consumption frequency, Food between meals, Physical activity frequency, Transportation used and log (Age).

Figure 8 – Regression Table of Model 2 Part A

	BMI		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	7.58	2.68 – 12.49	**0.002**
Gender [Female]	-2.25	-2.86 – -1.64	**<0.001**
Family.history.with.overweight [no]	-4.46	-5.33 – -3.59	**<0.001**
Fast.food.intake [no]	-2.06	-2.96 – -1.15	**<0.001**
Vegetable.consumption.freq [Never]	1.95	0.54 – 3.35	**0.007**
Vegetable.consumption.freq [Always]	1.58	0.97 – 2.20	**<0.001**
Food.between.meals [Always]	-3.69	-5.25 – -2.14	**<0.001**
Food.between.meals [Frequently]	-4.57	-5.65 – -3.49	**<0.001**
Food.between.meals [no]	-1.21	-3.03 – 0.61	0.192
Physical.activity.freq2-4 days	-0.20	-0.98 – 0.59	0.625
Physical.activity.freq4-5 days	-1.37	-2.56 – -0.17	**0.025**
Physical.activity.freq [Never]	0.89	0.19 – 1.58	**0.013**
Transportation.used [Personal Vehicle]	-2.28	-3.13 – -1.42	**<0.001**
Transportation.used [Walk/Bike]	-2.60	-4.13 – -1.06	**0.001**
log_Age	7.57	6.00 – 9.13	**<0.001**
Observations	785		
R2 / R2 adjusted	0.446 / 0.436		

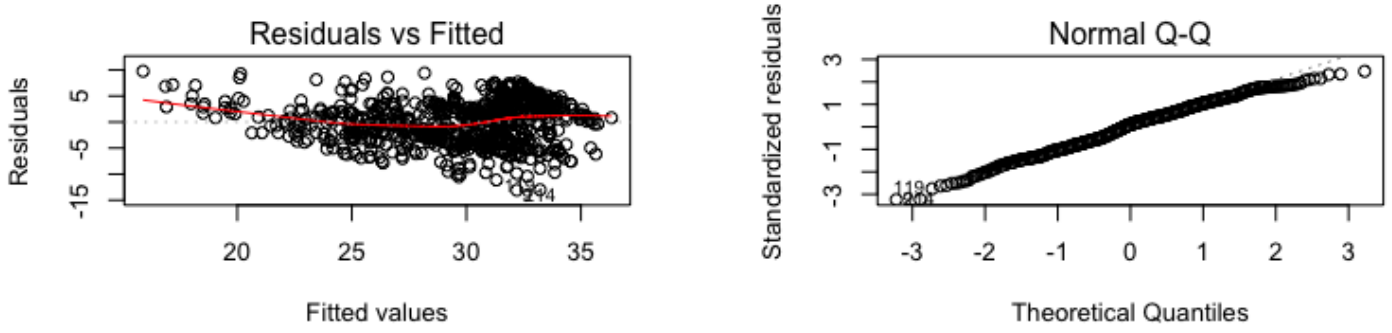
The intercept of Model 2 Part A is 7.58. This means that an individual has an expected BMI of 7.58 when all predictors equal their reference level and log (Age) equals 0.

Although Gender on its own had very little predictive power for BMI (Refer 2.3) , when compared with other predictors it is helpful to predict BMI. The coefficient estimate of Family history with overweight [no] was - 4.46 and this means that the predicted BMI of an individual who does not have a family history of overweight is 4.46 lower than someone that does. Further, if an individual frequently consumed food between meals, their predicted BMI would decrease by 4.57 compared to someone who consumed food between meals sometimes. The coefficient intercept of log (Age) is 7.57 and this means that a unit increase in log (Age), would lead to an increase in expected BMI by 7.57. This can also be expressed as a 1% increase in Age leads on average to an increase in the expected BMI by 0.0757.

The Residuals vs fitted plot of Model 2 Part A (Figure 9 – left) had a slight funnel shape, but this was not an issue. The Normal Q-Q plot (Figure 9 – right) lies along the x-y diagonal. However, slight kurtosis (deviation from the x-y diagonal) can be observed in the top right. Further, as the Cook's distance of all points are equal to or less than 0.02, there were no influential points.

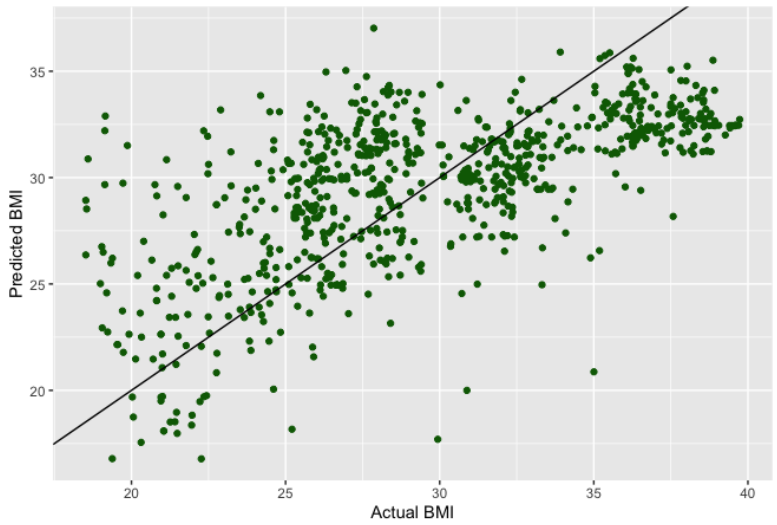
Multicollinearity was not an issue as the vif of the predictors was below 2.

Figure 9 – Residuals vs Fitted plot and Normal Q-Q Plot for Model 2 Part A



Then the testing set was used to plot the predicted value of BMI against the actual value of BMI (Refer Figure 10). Figure 10 reveals that the model predicts the BMI of individuals somewhat accurately.

Figure 10 – Actual vs Predicted BMI (Model 2 Part A)



3.1.3 MLR to predict BMI greater than 40 (Model 2 - Part B)

This model was constrained due to the small number of responses (sample size of 132) and the lack of variation in responses. Most individuals with a BMI exceeding 40 provided similar answers. Due to the homogeneity of responses, the following predictors were excluded from the model.

- Gender (1 Male and 131 Female)
- Family history with overweight (All individuals responded Yes)
- Fast food intake (1 No and 131 Yes)
- Vegetable consumption frequency (All individuals responded Always)
- Number of main meals (All individuals responded Three)
- Food between meals (1 Frequently and 131 Sometimes)
- Smoke (All individuals responded No)
- Calculate daily calories (All individuals responded No)
- Alcohol consumption (1 Never and 131 Sometimes)
- Transportation used (1 Personal Vehicle and 131 Public transport)

Further, the following levels of the respective predictors gained 0 responses

- The level ‘4-5 days’ for the Physical activity frequency predictor
- The level ‘More than 5 hours’ for the Time on tech predictor

Next, the base level of all categorical predictors was set to the most frequently occurring level (Modal level)

- More than 2 litres for Daily liquid intake
- Never for Physical activity frequency
- 3-5 Hours for Time on tech

A best subsets regression was run using the training data consisting of 132 observations. It was considered to fit the model using a larger training set. However, a lack of variation was observed even in a larger training set. Therefore, the training set consisting of 132 observations was used to model the data and the remaining 128 datapoints was used as the testing set. The regression model produced using the training set included Daily liquid intake and log (Age) as predictors.

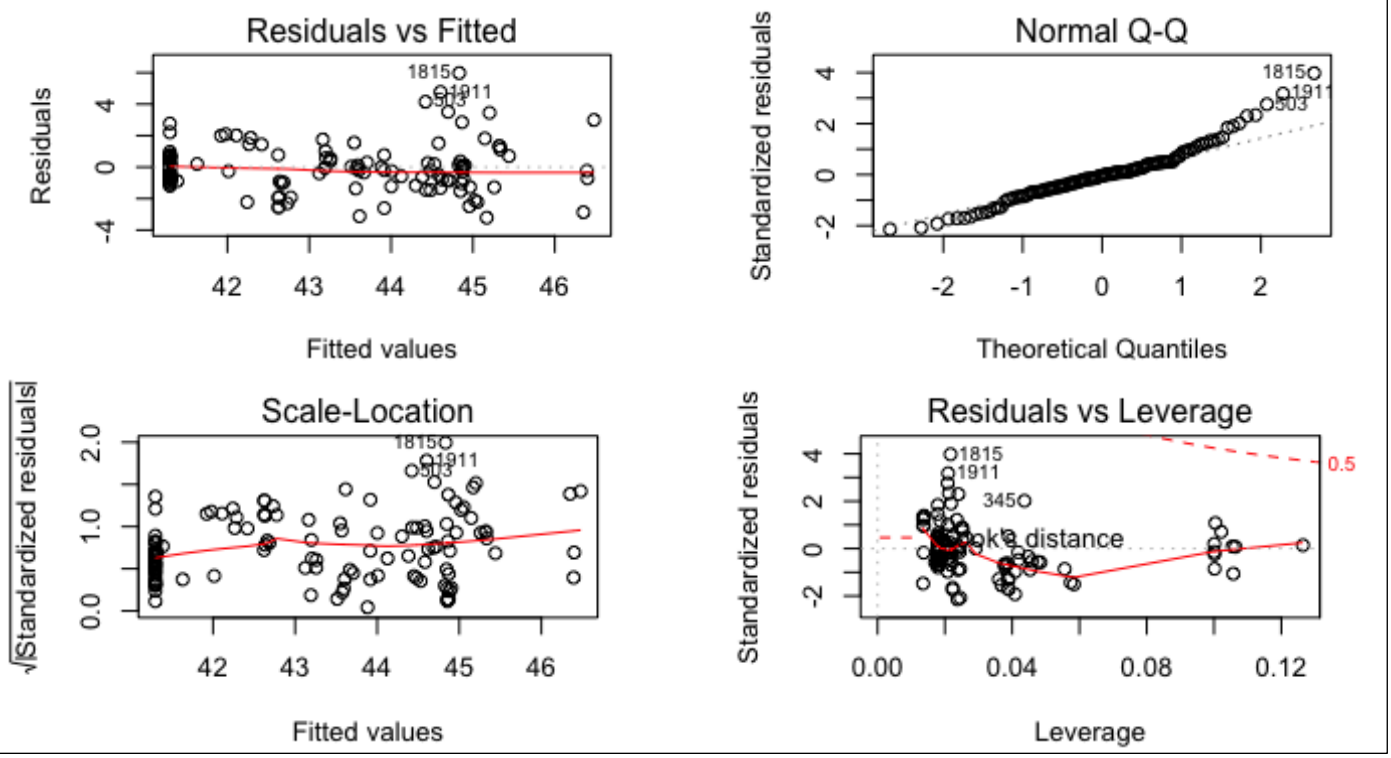
Figure 11 – Regression table of Model 2 Part B

	BMI		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	75.49	68.07 – 82.91	**<0.001**
Daily.liquid.intake1-2 litres	1.32	0.73 – 1.92	**<0.001**
Daily.liquid.intake [Less than a litre]	0.33	-0.72 – 1.39	0.533
log_Age	-10.49	-12.83 – -8.16	**<0.001**
Observations	132		
R2 / R2 adjusted	0.531 / 0.520		

The coefficient estimate of the intercept is 75.49. However, this is an unrealistic expected BMI. According to Model 2 Part B, the predicted BMI of individuals who consumed 1-2 litres of daily liquids or less than a litre of daily liquids was higher than those who consumed 2 or more litres of daily liquids (the base level). Surprisingly, a 1% increase in log (Age) led to a 0.1049 decrease in expected BMI.

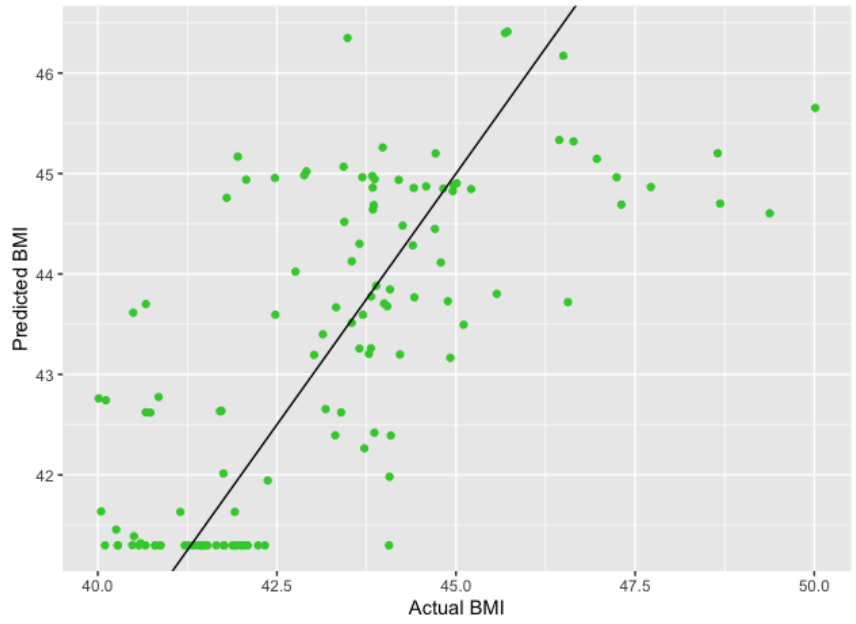
As the Residuals vs Fitted plot (Figure 12 – Top left) was pattern less, the residuals behaved as random noise. Normality of the residuals was observed as the Normal Q-Q plot (Figure 12 – Top right) lies along the x-y diagonal. However, the plot deviated from the diagonal on the far right. The Scale-Location plot (Figure 12 – Bottom left) shows a random scatter of points. This means that there are homogeneous variances among different observations and that the fitting is adequate. The Residuals vs Leverage plot (Figure 12 – Bottom right) reveals that data point 1815 was a candidate for an influential point. All data points except for 1815 had a Cook's distance less than 0.085. When the regression was rerun without this point, there was no change in the model. Therefore, 1815 was included in the model.

Figure 12 – Diagnostic plots of Model 2 Part B



Multicollinearity of the predictors was not an issue as the vif of both predictors was 1.18.

Figure 13 – Actual vs Predicted BMI (Model 2 Part B)



This model predicts the BMI over 40 well. A cluster of points on the bottom left (individuals with a BMI between 40 and 42.5) can be observed in Figure 13.

3.1.3 Multiple Linear Regression - Model 1 or 2?

Figure 14 – Comparing Model 1 and 2 side by side

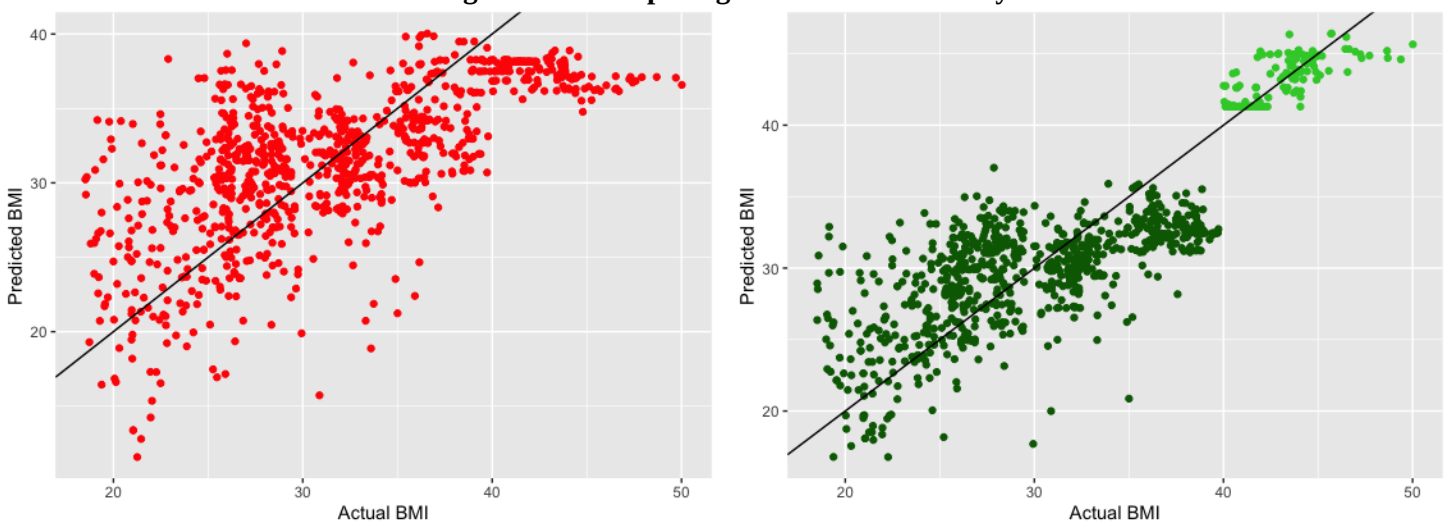


Figure 14 reveals that model 2 is clearly a better fit for the data. Further, the adjusted R^2 of model 1 was 0.49 whereas the adjusted R^2 value for model 2 Parts A and B was 0.44 and 0.52 respectively. This is clearly observed by the spread of the data in Figure 14. However, the R^2 of a model simply measures the spread of the data, and a low R^2 value does not mean zero correlation between the predictors and the dependent variable.

As Model 2 is better than Model 1, it was used understand the factors that affected obesity and provide tailored advice to obese individuals as to how they can decrease their BMI.

3.2 Trees

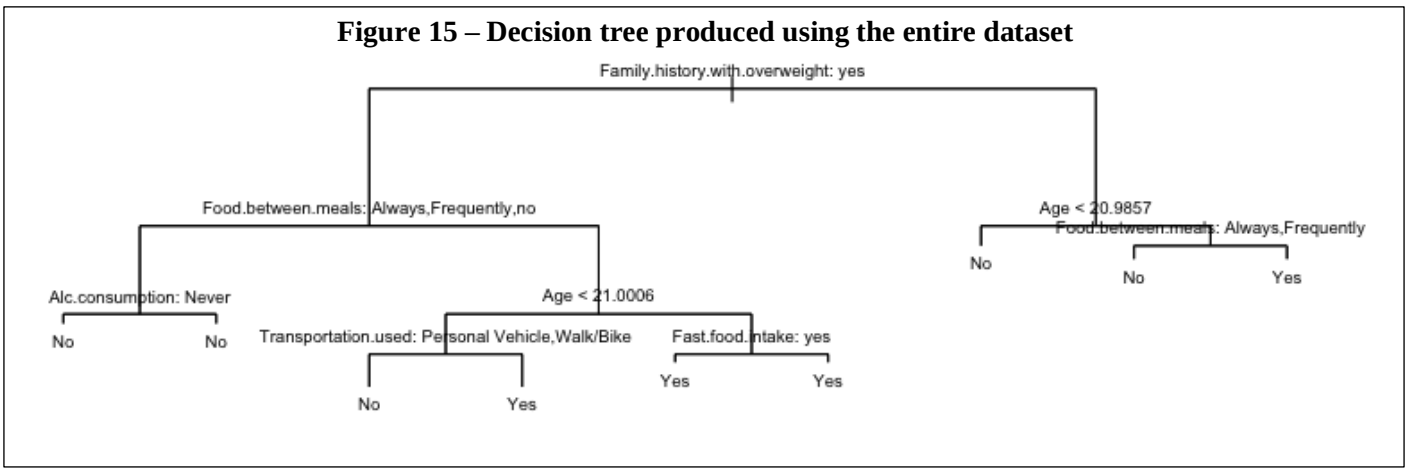
3.2.1 Decision Trees

A decision tree was used to identify the most influential factors that lead to a high BMI. As a decision tree produced a binary analysis of the situation, a new variable named 'High' was created ('Yes' indicates that an individual has a BMI greater than 25 and 'No' indicates that they have a BMI below 25). A combination of methods was used to obtain a simple but well-built tree that explained the outcome variable BMI.

Firstly, a simple classification tree based on the entire dataset was built. The minimum entropy (minimum deviance) measure was used to grow the tree. The variables actually used in the construction of the tree were:

1. Family history with overweight
2. Food between meals
3. Alcohol Consumption
4. Age
5. Transportation used
6. Fast food intake

Figure 15 – Decision tree produced using the entire dataset



The tree in Figure 15 had a total of 9 terminal nodes, the residual mean deviance for this tree was 0.6065 and the tree had a misclassification error rate of 11.42% (241 / 2111). Although the misclassification rate was 11.42%, the two types of errors (false positives and false negatives) have not been considered. These errors may result in incorrect advice being provided to individuals. For example, an individual who was wrongly categorised as overweight or obese may incorrectly be given advice to reduce their intake of food and change their lifestyle. This may negatively impact their health. On the other hand, a false negative will impact the individual who does not obtain the advice needed. Someone who is likely to have a high BMI may not be correctly identified at an early stage and this could greatly hinder their potential progress as it becomes harder to implement new habits as they get used to their usual activities.

As Family history of overweight individuals was the root node of the decision tree (Refer Figure 15), it was the most influential predictor. Other informative predictors are age and whether an individual has food between meals. As mentioned in this report, obesity may be caused by a combination of genetics and the environment that an individual

lives in. Further, factors such as Alcohol consumption, Transportation used, Fast food intake and Food between meals are predicted to lead to a high BMI.

An example inference from the tree in Figure 15 is that an individual who has overweight people in their family history, sometimes consumes food between meals, is over the age of 21 and consumes fast food regularly is predicted to have a high BMI.

After the tree was created, both CV and 10-fold CV were run to see if the tree could be pruned further. The criterion used to prune the tree was the misclassification rate. Both CV and 10-fold CV revealed that the tree cannot be pruned further and that the 9-node tree was the optimal one.

As it was necessary to assess the performance of the tree on new data, the observations were split into the training set (1056 observations) and the testing set (1055 observations). A second tree was the created using the training set and its performance was assessed using the training set. From Table 3, the misclassification rate involved when classifying new data from this tree was 10.9% ((60+55)/(222+55+60+718)). This tree accurately predicts whether data from a new set of individuals result in a high BMI in 89.1% ((222+718)/(222+55+60+718)) of cases.

Table 3: Confusion matrix for the decision tree

	Testing data: No	**Testing data: Yes**
Predicted: No	222	55
Predicted: Yes	60	718

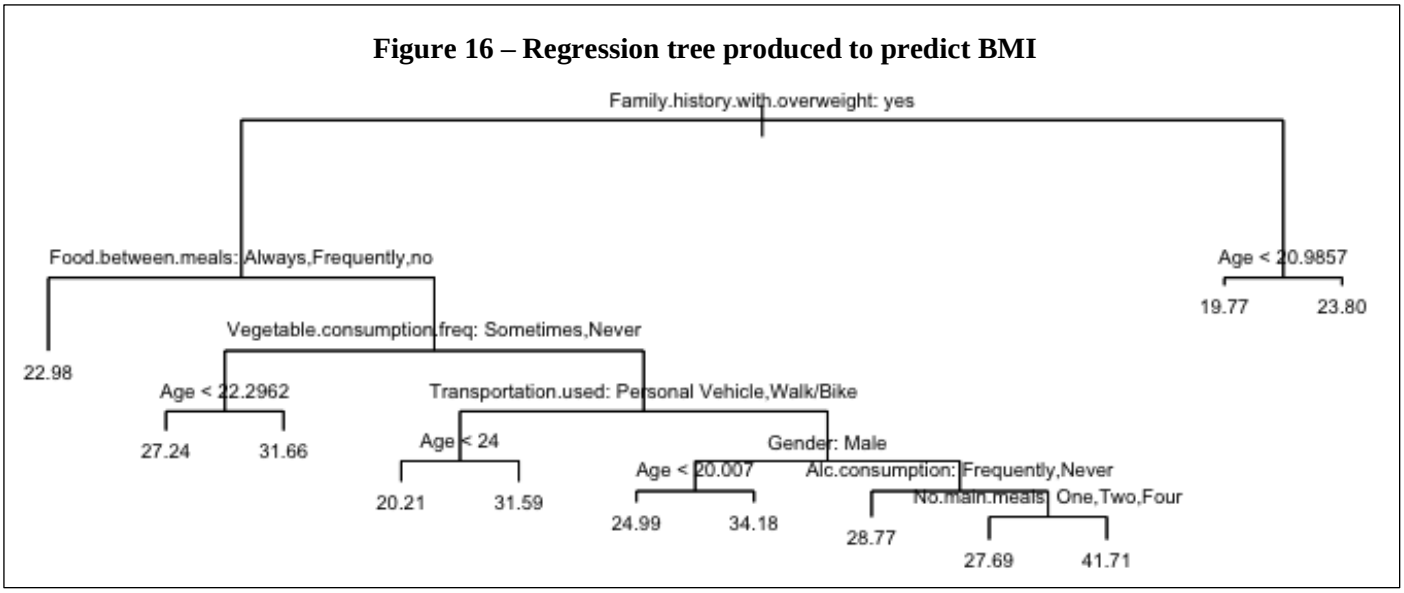
3.2.2 Regression tree

A new tree was created using BMI as the response variable to see how much each variable might impact a decrease or increase in BMI.

The tree was created using the training data and the remaining portion of data (testing data) was used to test the performance of the tree. The mean square predictive error was 20.8732 for the testing sample and this is slightly greater than the value for the training sample. The most important predictor was Family history with overweight.

An example inference from the tree in Figure 16 is that an individual who is 21 or older and has no family history of overweight individuals is predicted a BMI of 23.8.

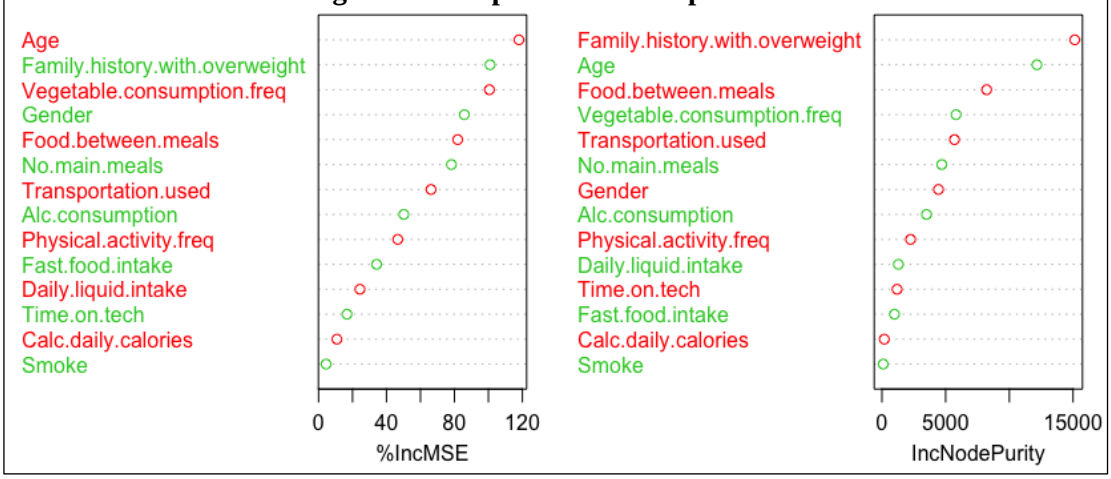
Figure 16 – Regression tree produced to predict BMI



3.2.3 Bagging

Bagging was performed to reduce the impact that the variance in the data had on the tree. Bagging is a bootstrap aggregation method which provides an average value of all the trees plotted using the training sets. Therefore, the variance is reduced from the original σ^2 to $\frac{\sigma^2}{n}$ and the larger the n (obtained by dividing the data into n smaller training sets), the lower the variance. The use of bagging reduced the mean Residual Sum of Squares (RSS) to 10.69873 and this was an improvement from the original tree. Further, testing it on the testing data gives a mean RSS of 11.02168.

Figure 17 – Importance of the predictors



The importance of each variable in the construction of the tree was also explored. This includes a measure of how the accuracy of the prediction improved upon the removal of each variable.

From Figure 17, Family history with overweight was the most significant predictor. Age and gender are other significant variables.

3.2.4 Random forest

Random forest is similar to bagging because it reduces the variance. However, at each split the variables that could be included are reduced. A standard choice for the variables that could be included is usually the square root of the number of predictors in the original sample data. If there is a dominant variable which is likely to enter all trees, then this would make all trees highly correlate with each other and affect the final averaged tree. Random forest reduces the correlation

by removing these dominant variables from being used at every split of the tree. Running the tree using random forest reduces the number of variables tried at each split to 7 and as a result the mean of squared residuals reduces to 10.04003. This is lower than when bagging was used.

3.3 Logistic regression

A logistic regression was used to determine the predictors that lead to an increased probability of having a high BMI.

The regression model was created using a training set (including 1056 data points). The dependent variable was the binary outcome 'High' (Refer 3.2.1). Each time the generalised linear model was run, the predictor with the highest p - value was removed as it was the least significant predictor. Therefore, Calculate daily calories, Vegetable consumption frequency, Time spent on technology, Fast-food intake and Gender were removed respectively. The final model using the entire dataset contained only the predictors that were significant at the 5% level. It included Age, Family history with overweight, No of main meals, Food between meals, Smoke, Daily liquid intake, Physical activity frequency, Alcohol consumption and Transportation used.

Figure 18 – Regression table of the logistic regression

		High	
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.08	0.02 – 0.29	**<0.001**
Age	1.27	1.20 – 1.35	**<0.001**
Family.history.with.overweight [no]	0.07	0.04 – 0.11	**<0.001**
No.main.meals [One]	1.11	0.64 – 1.96	0.709
No.main.meals [Two]	5.86	1.82 – 24.15	**0.007**
No.main.meals [Four]	0.36	0.16 – 0.79	**0.011**
Food.between.meals [Always]	0.16	0.05 – 0.48	**0.001**
Food.between.meals [Frequently]	0.04	0.02 – 0.07	**<0.001**
Food.between.meals [no]	1.04	0.36 – 3.25	0.938
Smoke [yes]	0.17	0.05 – 0.63	**0.007**
Daily.liquid.intake [Less than a litre]	1.14	0.68 – 1.95	0.617
Daily.liquid.intake [More than 2 litres]	2.04	1.19 – 3.57	**0.010**
Physical.activity.freq2-4 days	0.48	0.28 – 0.81	**0.007**
Physical.activity.freq4-5 days	0.37	0.17 – 0.83	**0.015**
Physical.activity.freq [Never]	0.73	0.43 – 1.24	0.241
Alc.consumption [Frequently]	1.77	0.58 – 5.75	0.330
Alc.consumption [Never]	0.46	0.29 – 0.72	**0.001**
Transportation.used [Personal Vehicle]	0.21	0.11 – 0.40	**<0.001**
Transportation.used [Walk/Bike]	0.17	0.06 – 0.50	**0.001**
Observations	1056		

If the odds ratio of a level of a predictor is less than one, this level decreases an individual's likelihood of having a high BMI when compared to the base level. Similarly, if the odds ratio is greater than 1, the level increases the likelihood of having a higher BMI compared to the base level.

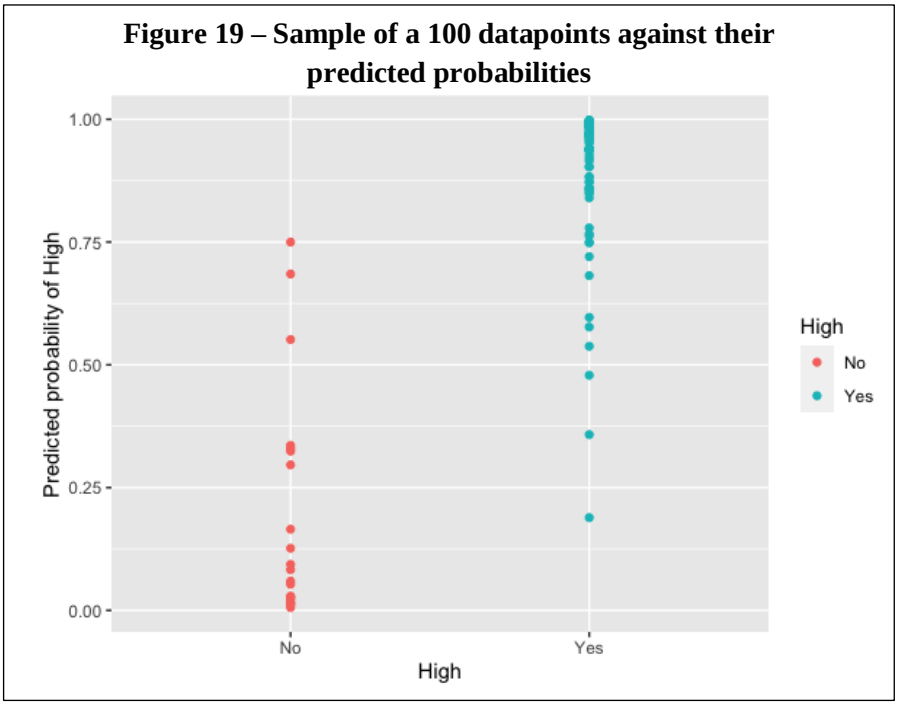
The testing set consisted of 1055 data points. The confusion matrix in Table 4 reveals how well the model predicts for the testing set.

Table 4 Confusion matrix for the logistic regression

	Testing data: No	Testing data: Yes
Predicted: No	200	37
Predicted: Yes	82	736

This shows us that out of all the observations in the testing set, 95.2% (736/(37+736)) of people with a high BMI were predicted correctly and 70.9%(200/(200+82)) of those who do not have a high BMI were predicted correctly.

Figure 19 – Sample of a 100 datapoints against their predicted probabilities



To visualize this, a sample of 100 random points have been plotted against their probabilities. Anything above 0.5 was predicted as a High BMI, while everything below is predicted to not be a high BMI. The bulk of the data belonging to those who have a high BMI are correctly predicted, with the data forming a tight band between 0.8 and 1. There are a several points between 0.5 and 0.75 and three points below 0.5. The values of individuals who do not have a high BMI are spread out a bit more. The datapoints form a loose group between 0 and 0.125. Three points are predicted probabilities exceeding 0.5 and thus these are incorrectly predicted. These visualizations are supported by Table 4.

4. Conclusion

4.1 Limitations associated with the data

This report was based on data collected from individuals residing in Mexico, Peru and Colombia. However, this dataset did not include the country that the individual was from. Therefore, the levels of obesity and factors contributing to obesity in the respective countries could not be compared.

There is a possibility that responders did not give completely honest answers due to the sensitivity of the questions asked. Moreover, 77% of the data was simulated. Therefore, these data may not be reflective of individual's characteristics.

Information regarding physiological factors such as level of stress, physical and mental illnesses and eating disorders were not included in the dataset. As these factors may also lead to overeating, knowledge of them will help to provide accurate and helpful advice to obese individuals. Further, an individual may not have similar lifestyle and dietary habits throughout their life. Therefore, advice provided may not be relevant in the long run.

4.2 Findings

Understanding the factors associated with increasing levels of obesity is a crucial step in being able to tackle the health issues related to it.

Based on the statistical learning procedures used, the main factors that are predicted to contribute to obesity are as follows.

Factors that an individual cannot change but should be aware of:

- **Age**: As people get older their expected BMI increases. However, this relationship was not observed in the multiple linear regression predicting BMI exceeding 40 (Model 2 Part B).
- **Family history**: People with a family history of BMI exceeding 25 are more prone to being overweight.
- **Gender**: Females are more likely to have a higher BMI than males.

Factors that an individual can change:

- **Food between meals**: Surprisingly, those who frequently consume food between meals tend to have a lower expected BMI than those who consume food between meals sometimes.
- **Fast food intake**: As expected, individuals who have an increased consumption of fast food are predicted a higher BMI than those who do not.
- **Physical activity frequency**: Increased physical activity led to a decrease in BMI.
- **Transportation used**: Individuals who walked/biked had a lower BMI than those who used public transportation.

4.2.1 Findings from the MLR used to predict BMI (Model 2)

BMI under 40 (Model 2 Part A): The factors that greatly influenced an individual's BMI were Gender, Age, Family history of high BMI, Fast food intake, Food between meals and Transportation used. Females had a lower BMI than their male counterparts. The older an individual is, the greater the predicted BMI. As expected, family history of high BMI and fast-food intake positively correlated with BMI. Consuming food between meals frequently or always reduced an individual's BMI when compared to eating food between meals sometimes. Those who walk/bike to work have a

lower expected BMI than others, closely followed by those with a personal vehicle; those who take public transport have the highest predicted BMI.

BMI over 40 (Model 2 Part B): The model reveals that factors such as Daily liquid intake and Age are closely related to BMI. Those who drink 1-2 litres of daily liquid experience an increase in the predicted BMI compared to individuals who drink more than 2 litres of daily liquid. An increase in the age of an individual will cause a decrease in the expected BMI of the individual.

4.2.2 Findings from the Tree

The main factor associated with obesity was Family history with overweight. If an individual had a family history of overweight, they were more likely to have a high BMI (BMI exceeding 25). This may be because of two reasons. Firstly, obesity may be genetic. Secondly, an individual may have identical or similar dietary and eating habits to the rest of their family.

Age was another factor that impacted whether an individual had a high BMI. The positive correlation between the two may be because the Basal Metabolic Rate (BMR) slows down as an individual gets older. Food between meals was another factor that directly contributed to a high BMI. However, this depends on the type of food consumed (is the food high in caloric intake) and the quantity of food consumed.

Transportation used and Alcohol consumption were other predictors that were used in the construction of the trees.

4.2.3 Findings from the GLM

The GLM suggests that the following predictors have the most impact on the probability of an individual having a BMI greater than 25 (High BMI):

- Daily liquid intake - A daily liquid intake of more than two litres increases an individual's odds of having a high BMI the most.
- Number of main meals - Those who consume two meals have a higher probability of being overweight or obese than those who consume three meals. Having four meals a day decreases an individual's probability of having a high BMI.
- Food between meals – Consuming food between meals always or frequently decreases an individual's probability of having a high BMI.
- Family history of overweight – Having a family history of overweight increases an individual's probability of having a high BMI.
- Transportation used – Using a personal vehicle or walking/biking decreases an individual's probability of having a high BMI compared to public transportation.
- Physical activity frequency – The more an individual exercises, the lower an individual's probability of having a high BMI.

- Alcohol consumption – The more alcohol an individual consumes the higher an individual's probability of having a high BMI.

4.2.4 Relevance of findings to the practical problem

Our findings are relevant in assessing the factors related to obesity of individuals in Mexico, Colombia and Peru. However, there is a possibility that these findings cannot be generalised to the rest of the world.

4.3 Improvements

The data included individuals aged 14 and over. However, there has been increased concern regarding childhood obesity in the media and health profession. Therefore, there is scope for an even wider discussion if the existing data is combined with that of children. Then this dataset can be used to investigate the factors affecting obesity and compare the differences. An interesting approach would be to see if habits developed in childhood affect their health later in life.

If the data set included the country that the individual was from, the similarities and differences between the factors that affected obesity in these three countries could have been compared.

4.4 Advice provided

Findings from the individual models can be combined to provide tailored advice to patients. It is important to understand an individual's current lifestyle and tailor the recommendations accordingly.

Factors such as age and family history contribute to a high BMI. Therefore, it is important for individuals to routinely engage in some form of physical activity to maintain their weight and strength. Further, the Physical activity frequency of an individual was a factor that negatively correlated with BMI. If an individual had no or inadequate physical activity the reason for this should be understood. If the reason was that individuals was busy with work, they could be advised to leverage on their method of transport to compensate for the lack of physical exercise. They could change their method of transport by substituting the use of a car or other personal vehicle (which requires limited movement) to either walking or cycling. However, the practicality of this should be considered.

Consuming food between meals was another factor that led to overweight and obesity. Therefore, individuals should be advised to pre-plan meals and frequently consume healthy snacks.

Regular fast-food consumption led to a high BMI. Therefore, it was important to advise individuals to stop or heavily limit their fast-food consumption.

Transportation used was another factor that affected BMI. As individuals who walked/biked had a lower BMI than the rest, all individuals could be advised to look at the practicality of incorporating this as a form of cardio in their daily life.

However, the above advice was provided solely based on the findings. Therefore, this advice should be used in conjunction with other medical advice.

4.5 Key Takeaways

The aim was to use statistical procedures to identify and analyse the factors that contributed towards obesity of individuals and use this information to provide tailored advice to individuals to combat obesity. The analysis helped identify the factors which greatly influenced a high BMI.

The findings help identify the main factors which are associated with a raised BMI within individuals. This helps to target resources, efforts and funds towards the key areas of concern regarding obesity. to maximise the outcome of any steps taken to help individuals gain control of their health. While the issues and the level of severity varies between countries, obesity has been identified as a global issue. The models and our findings are likely to stay relevant in the future as multiple studies show similar connections between the identified factors and BMI. However, there may be some changes to their significance as the environment changes. As education and work become more virtual, time on technology may become more influential than transportation used.

5. Bibliography

Body Mass Index (BMI). WHO global health observatory. Available at https://www.who.int/data/gho/data/themes/theme-details/GHO/body-mass-index-(bmi).

Causes: Vitamin B12 or folate deficiency anaemia. NHS. Available at https://www.nhs.uk/conditions/vitamin-b12-or-folate-deficiency-anaemia/causes/.

Dhurandhar, E.J., Keith, S.W. The aetiology of obesity beyond eating more and exercising less. Best Practice & Research Clinical Gastroenterology, Volume 28, Issue 4, August 2014, Pages 533-544. DOI: 10.1016/j.bpg.2014.07.001

Elks, C.E., den Hoed, M., Zhao, J.H., Sharp, S.J., Wareham, N.J., Loos, R.J.F., Ong, K.K. (2012). Variability in the heritability of body mass index: a systematic review and meta-regression. Frontiers in Endocrinology, February 2012; **3**: 29. DOI: 10.3389/fendo.2012.00029

Hruby, A., Manson, J.E., Qi, L., Malik, V.S., Rimm, E.B., Sun, Q., Willett, W.C., Hu, F.B. Determinants and consequences of obesity. A Public Health of Consequence: Review of the September 2016 Issue of the American Journal of Public Health, 2016;106(9):1656–1662. DOI: 10.2105/AJPH.2016.303326

Mandal, Ananya. (February 2019). Medical news. Obesity and fast foods. Available at https://www.news-medical.net/health/Obesity-and-Fast-Food.aspx.

Obesity. World Health Organisation (WHO). Available at https://www.who.int/news-room/facts-in-pictures/detail/6-facts-on-obesity.

Obesity Statistics. The European Association for the Study of Obesity (EASO). Available at https://easo.org/media-portal/statistics/.

Qi, Q., Chu, A.Y., Kang, J.H., et al. Fried food consumption, genetic risk, and body mass index: gene–diet interaction analysis in three US cohort studies. British Medical Journal, 2014;348:g1610. DOI: https://doi.org/10.1136/bmj.g1610

Qi, Q., Chu, A.Y., Kang J.H., et al. Sugar-sweetened beverages and genetic risk of obesity. The New England Journal of Medicine, 2012;367(15):1387–1396. DOI: 10.1056/NEJMoa1203039.

Stop smoking without putting on weight - quit smoking, NHS. Available at

https://www.nhs.uk/live-well/quit-smoking/stop-smoking-without-putting-on-weight/#:~:text=Smoking%20can%20suppress%20your%20appetite,action%20of%20smoking%20with%20snacking.

Taylor, Rachael. (June 2011). The Science Learning Hub. Obesity: Genetic or environmental. Available at https://www.sciencelearn.org.nz/resources/203-obesity-genetic-or-environmental.

Valera, B., Sohani, Z., Rana, A., Poirier, P., Anand, S.S**.** The ethnoepidemiology of obesity. Canadian Journal of Cardiology, Volume 31, Issue 2, February 2015, Pages 131-141**.** DOI: https://doi.org/10.1016/j.cjca.2014.10.005

Qasim, A., Turcotte, M., de Souza, R.J., Samaan, M.C., Champredon, D., Dushoff, J., Speakman, J.R., Mevre, D. On the origin of obesity: identifying the biological, environmental and cultural drivers of genetic risk among human populations. Obesity Reviews, Volume 19, Issue 2, February 2018, Pages 121-149 (2). DOI: https://doi.org/10.1111/obr.12625

6. Appendix

6.1 Data source

The dataset used for the purpose of this report was the Estimation of obesity levels based on eating habits and physical condition. This was obtained from the UCI machine learning repository and is available via the following link: https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+

6.2 The questions and possible answers of the online survey used to collect the data

More information regarding the dataset can found in the following paper.

Mendoza-Palechor,F., de la Hoz Manotas, A. Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. Journal data in brief, Volume 25, August 2019, 104344. DOI: 10.1016/j.dib.2019.104344.

This included information regarding the survey questions used to create the dataset and variables included in the dataset.

The questions used in the survey are as follows.

1. What is your gender? - Female, Male
2. What is your age? - Numeric value
3. What is your height? - Numeric value in metres
4. What is your weight? - Numeric value in kilograms
5. Has a family member suffered or suffers from overweight? Yes, No
6. Do you eat high caloric food frequently? Yes, No
7. Do you usually eat vegetables in your meals? Never, Sometimes, Always
8. How many main meals do you have daily? Between 1 and 2, 3, More than 3
9. Do you eat any food between meals? No, Sometimes, Frequently, Always
10. Do you smoke? Yes, No
11. How much water do you drink daily? Less than a litre, Between 1 and 2 litres, More than 2 litres
12. Do you monitor the calories you eat daily? Yes, No
13. How often do you have physical activity? I do not have, 1 or 2 days, 2 or 4 days, 4 or 5 days

14. How much time do you use technological devices such as cell phone, videogames, television, computer, and others? 0–2 hours, 3–5 hours, More than 5 hours
15. how often do you drink alcohol? I do not drink, Sometimes, Frequently, Always
16. Which transportation do you usually use? Automobile, Motorbike, Bike, Public Transportation, Walking