Analysis on Socioeconomic Status, Diet and Obesity

Sung Ho Kim, Shirley Ma, Sybilla Levenston, Nina Weiss, Yueyi Liu

Contents

1 Executive Summary

The relationship between an individual's income and their diet has been well documented in research literature. Australia currently faces the rising problem of obesity, where almost 36 percent of adult Australians recorded being overweight and more than 1 in 4 Austrlians recorded being obese (Foundation 2017). While the statistics may be grave, there is much research being done into solutions that are easily implementable. The focus of this report is on one such solution, that being increasing the awareness of high levels of dietary fibre in preventing obesity and the diseases it entails (Paris Michel-Ange) 2014).

In order to deconstruct this problem, we have first investigated correlations between disease and income. The expectation was that whilst disease status and income might have a correlation, this correlation would be quite weak. The underlying reason lay with an individual's income having a crucial role in determining the foods they purchased. The findings matched what we had suspected; when applying χ^2 tests it was found that two of the three obesity-related diseases (high cholesterol and hypertension) shared a dependance structure with income. However, Cramer's V for all three comparisons culminated with the same conclusion which was that the association between income and disease was weak.

The conclusions shed light on other factors such as diet and lifestyle habits which are majorly influenced by income levels. kNN and Linear Discriminant Analysis were two methods implemented in the construction of a classifier attempting to predict whether an individual would suffer from high cholesterol. Factors of interest included: weekly exercise duration, BMI levels, waist circumference and percentage of daily energy intake derived from trans and saturated fats. Both methods yielded a cross-validation error of 11.4 percent for males and 9.84 percent for females. Interest then turned on the importance of fibre in promoting health and multivariate linear regression was performed examining the trends of dietary fibre consumption within unhealthy individuals. The models were constructed based on seven dietary elements deemed important by our friends studying nutrition and in almost all instances, a poor state of health was characterised heavily by a decline in fibre consumption. Principal Component Analysis (PCA) was implemented to further justify

the importance of fibre, with the conclusion that fibre played a major role in each of the first principal components.

The last section explores the correlation between fibre intake and disease status, with CART classifiers being implemented on the datasets. The datasets were stratified accordingly, firstly by gender, secondly by income brackets. The CART analysis provided a readily interpretable visual analysis of the key risk factors associated with high cholesterol, high blood sugar levels and hypertension.

2 Motivation and Agenda

The obesity epidemic in Australia can be seen as a wicked problem, many causes with no simple solution. A strong aspect to explore in the amelioration of obesity, is through dietary intervention. We have chosen to look at the role that fibre plays in mediating or even preventing metabolic disease. Dietary fibre is a nutrient that cannot be entirely digested or absorbed. Dietary fibre can improve chronic disease by regulating satiety, blood sugar and bowel health. The consumption of 5 servings of whole grains per day is useful in lowering the risk of obesity by 10% in men and 4% in women (Van de Vijver and Goldbohm 2009). Controlling satiety is an important factor in regulating body weight, which can be suppressed by the viscous texture of soluble fibre in the gut and therefore increase the time of intestinal digestion and absorption from other nutrients (Slavin 2007). In terms of cardiovascular diseases, dietary fibre intake can positively affect blood pressure, blood cholesterol and blood glucose levels. The experiments that has been conducted with long term intake of 10g of dietary fibre supplement per day shows a decline of approximately 1-2mmHg in both systolic and diastolic pressure (Delzenne 2005). Another 3 day study suggests that LDL-cholesterol and fasting glucose level would decrease by 12% in healthy individuals when comparing low (10g/day) and high (30g/day) fibre intake diet (Aller 2004). Additionally, the fermentation of dietary fibre in the colon is beneficial to gut bacteria and healthiness of colon cells for preventing colorectal cancer. Additionally, the risk of developing type 2 diabetes would be significantly reduced by an increased intake of dietary fibre. (Buttriss 2008). Thus, the effects of dietary fibre on obesity- and chronic-diseases-related variables is the focus of this project.

3 Research Questions and Statistical Formulation

1. Whether income has any impact on an individual's chance to develop diseases related to obesity. How strong is this impact?

2. What are unhealthy males and females doing wrong when it comes to their diet? We would like to investigate this in the context of their income, so is the trend the same across poor, middle class and wealthy people? What is the trend and overall importance of fibre to a healthy diet?

3. We would like to also investigate other factors, on top of fibre which could prevent high cholesterol, high blood sugar levels and hypertension.

From a statistical viewpoint, the aforementioned questions could be reinterpreted and approached as follows:
1. Perform tests of independence between income and the three chosen diseases. Determine the strength of their association.
2. Take key components of daily diet as predictor variables and construct an appropriate model which underlines the trends present within unhealthy individuals. The notion of unhealthy is henceforth pinned to a BMI classification of overweight and higher.
3. Develop a classifier which incorporates fibre as well as factors relating to the disease for which we are trying to classify. Draw conclusions as to what the key risk factors are and if they are preventable in any way.

4 Correlations and Classifiers: Income and Disease

4.1 Question Focus:

The first question attempts to determine whether a correlation is present between an individual's income status and whether they suffer from obesity-related diseases. The diseases of interest are high cholesterol levels (HCHOLBC), high sugar levels (HSUGBC) and hypertensive disease (HYPBC).

An initial check for independence has been accomplished through Pearson's χ^2 test.

```
chi1<-chisq.test(table1)
chi1
```

```
##
##  Pearson's Chi-squared test
##
## data:  table1
## X-squared = 95.042, df = 2, p-value < 2.2e-16
```

The χ^2 test for independence yields a p-value of $2.2 * 10^{-16}$. We can conclude from this p-value that there is significant evidence to suggest the existence of some kind of dependancy structure between an individual's income levels and whether they suffer from high cholesterol levels.

```
chi2<-chisq.test(table2)
chi2
```

```
##
##  Pearson's Chi-squared test
##
## data:  table2
## X-squared = 6.0369, df = 2, p-value = 0.04888
```

The χ^2 test here yields a p-value of 0.06 (≈ 0.05) and we conclude there is insufficient evidence to suggest a dependency between socioeconomic status and high sugar levels.

```
chi3<-chisq.test(table3)
chi3
```

```
##
##  Pearson's Chi-squared test
##
## data:  table3
## X-squared = 195.05, df = 2, p-value < 2.2e-16
```

The χ^2 test for income vs. hypertension yields a p-value $< 2.2e-16$. From this p-value we draw the conclusion that there exists a dependancy structure between income and whether the individual suffers from hypertension.

4.1.1 Cramer's V: A Measure of Association

Cramer's V was then implemented in order to give a better indication about the strengths of association.

Comparison	Cramer's V
HCHOLBC	0.126
HSUGBC	0.032
HYPBC	0.180

The Cramer's V for INCDEC vs. HSUGBC is extremely low, suggesting there is a negligible association between these two variables. This agrees with the p-value obtained in the corresponding χ^2 test. While the other two coefficients are substantially larger, they still suggest extremely weak associations.

4.1.2 kNN and LDA Classifiers for High Cholesterol Status:

The weak associations give rise for taking the report in another direction. We have substantial evidence to show that income clearly plays a role in determining disease status of an individual, however insufficient that interaction may be. We can begin with a rudimentary implementation of kNN and LDA classifiers for one of the diseases. Discussion with nutrition students elucidated that high cholesterol status is majorly affected by the constituents of one's diet. Other risk factors included obesity, large waist circumference and lack of exercise, which were all taken into account upon building the classifier. The aim of this section is thus being able to create a classifier which can predict what kind of individuals would suffer from high cholesterol.

We are interested in comparing kNN with LDA performances based off of CV error percentages.

```
b
```

```
##   k  Gender kNN CV Error
## A 13 male   0.1138
## B 15 female 0.09838
```

```
as.table(a)
```

```
##   Gender LDA CV Error
## A Male   0.1138
## B Female 0.09838
```

The conclusions drawn from LDA are: ***Males***
1. A higher intake of saturated and trans fats lead to a higher probability of high cholesterol.
2. A sedentary lifestyle raises chances of high cholesterol levels. We can infer is that individuals who live a highly sedentary lifestyle are also more likely to live off of an unhealthy diet.
3. As an individual's BMI goes up, this matches with an increased chance of suffering from high cholesterol.

Females
1. An increased intake of saturated and trans fats leads to a higher probability of the female suffering from high cholesterol.
2. As a female's BMI increases, so does their chances of suffering from high cholesterol.

Furthermore, the CV errors for these are 11.4 percent for males and 9.84 percent for females which are the same percentages derived from choosing the optimal amount of clusters in kNN analysis. However, we would like to draw attention to the counterintuitive sign for the coefficient determining weekly exercise. LDA analysis concludes that an individual, both male and female is more likely to suffer from high cholesterol levels if they exercise more often. Upon raising the issue with the nutrition students, it was posited that high cholesterol status is predominantly controlled by diet and not lifestyle choices or genetics and thus this issue could be disregarded.

5 Dietary trends within unhealthy individuals

5.1 Question Focus:

The second component of the report draws upon the statistical techniques offered to purely numerical data. The focus of this question is to draw conclusions from Multivariate Linear Regression and Principal Component Analysis in highlighting dietary trends of individuals with low income. We will then attempt to compare their diet with people in higher income brackets. In preparation for the aforementioned analysis, we

have taken the liberty with first stratifying the sample by gender. Consecutive stratification by income level is further performed. The second stratification step is quintessential, providing the means for adequately satisfying the normality assumptions underlying the regression analysis. It would like to be stated that it was difficult to normalise all the residuals, even after performing Yeo-Johnson transformations where appropriate on the response variable.

Collaboration with NUTM students yielded that:
1. Daily Fibre Intake (FIBRET1)
2. Daily Sugar Intake (SUGART1)
3. Daily Saturated Fat Intake (SATFATT1)
4. Daily Polyunsaturated Fat Intake (PUFATT1)
5. Daily Monounsaturated Fat Intake (MUFATT1)
6. Daily Fruit Intake (FRUIT1N)
7. Daily Vegetable and Legume Intake (VEGLEG1N)

are important predictor variables in determining state of health.

We adopted Green's (ResearchGate 2016) rule of thumb for minimum sample size; since we are testing the overall model, each of the data sets having regression performed onto them should have a minimum of 106 observations for an acceptable sample size. The conclusion that we draw is that there are a sufficient amount of data points which can guarantee to some degree, the statistical power desired for this regression for the lowest income class. Unfortunately, we are unable to extend this proposition to the other male classes as their sample size, while larger than the minimum required isn't enough to guarantee strong statistical power. The next step will be to conduct some initial diagnostic plots, to verify that each of the data sets satisfy the assumptions of normality required in order to perform regression.

```
#Residual Test for Non-transformed Dataset
shapiro.test(lm.m.p$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm.m.p$residuals
## W = 0.8914, p-value < 2.2e-16
```

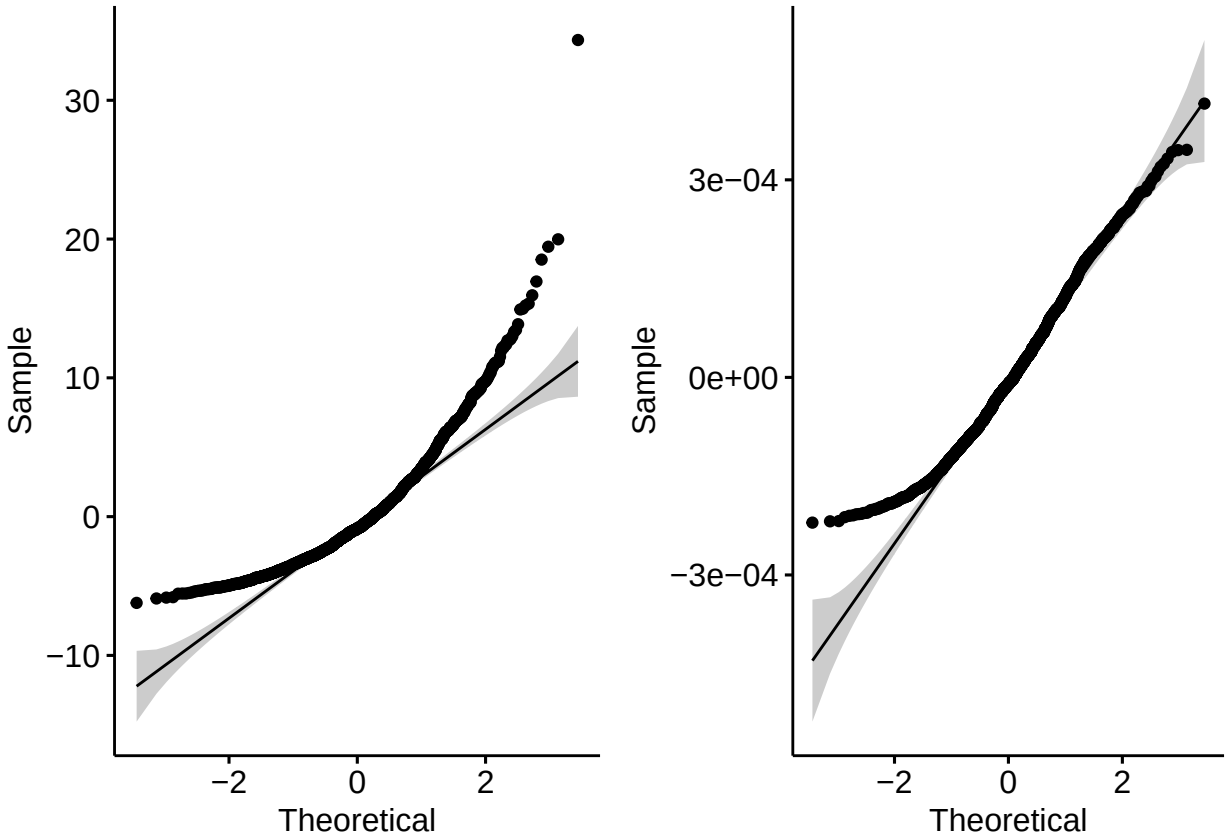
```
##
## Attaching package: 'huxtable'
```

```
## The following object is masked from 'package:ggpubr':
##
##     font
```

```
##
## Attaching package: 'kableExtra'
```

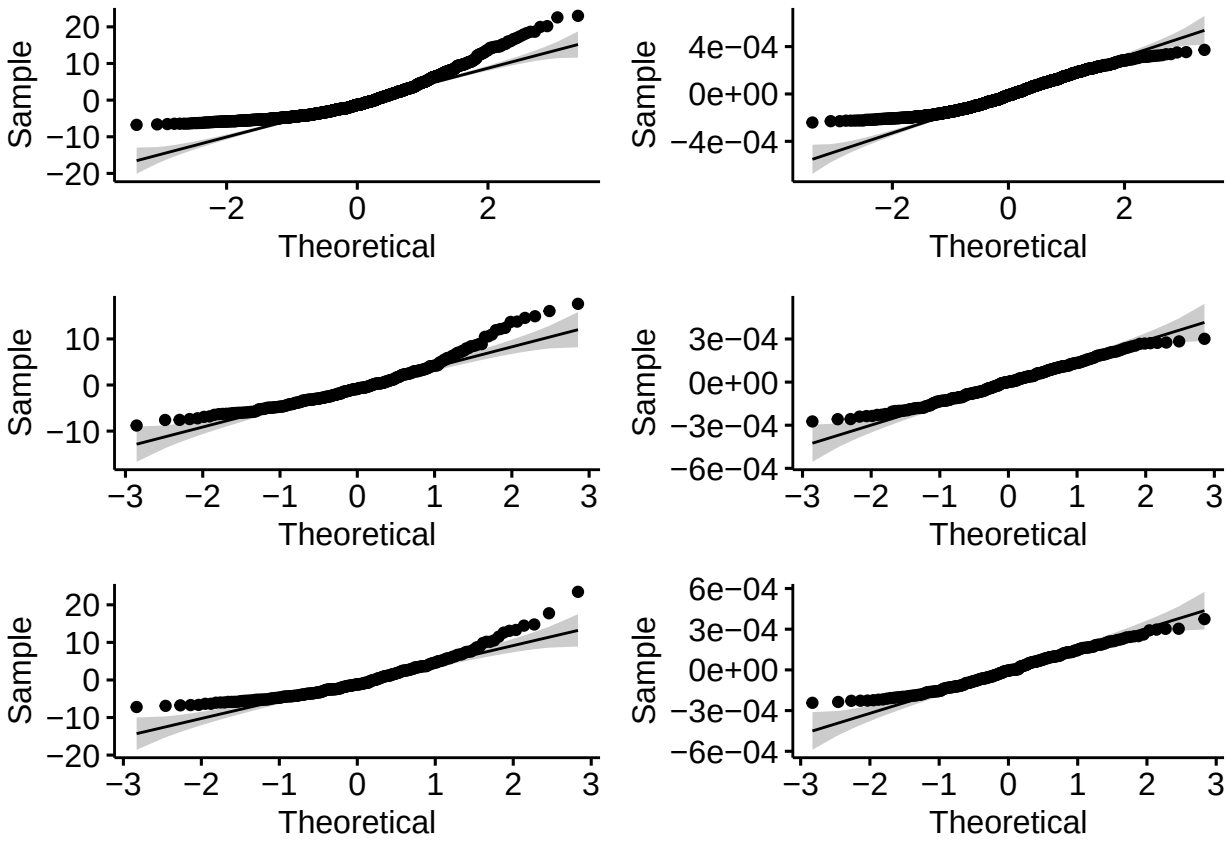
```
## The following object is masked from 'package:huxtable':
##
##     add_footnote
```

Coefficients	Normal	Transformed
Intercept	30.5	0.499
Fibre	-0.0525	-1.51e-06
Sugar	0.00346	7.83e-08
Sat Fat	0.00029	4.44e-08
Polyunsat Fat	-0.0115	-3.17e-07
Monounsat Fat	0.00782	2.41e-07
Fruit	-0.0494	-1.35e-06
Veg	0.0117	4.19e-07



The qq-plot and Shapiro-Wilk's normality test clearly indicate that the residuals do not satisfy normality. A solution is the implementation of a Yeo-Johnson transformation onto the response variable BMISC in an attempt to centre most of the residuals. It can not be guaranteed that the transformation will be able to cater for the extreme residuals, but we shall proceeed anyway. We note that the Yeo-Johnson transformation does an effective job at normalising the residuals at the upper tail region but is lacking with the lower tailed residuals. More importantly, a male within a lower socioeconomic bracket suffers from decreased fibre and fruit consumption and an increase in saturated fat intake and sugar. Possible reasons include these foods being more affordable and energy dense options that are nutritionally deficient. Being in a lower SES bracket, the individual's choices would be predominantly spearheaded by costs and familiarity thus leading to a fibre-lacking diet.

Shapiro.p.Value	**Normal**	**Transformed**
Females: Lowest Income Class	5.8e-30	5.84e-17
Females: Middle Income Class	9.53e-09	0.0453
Females: Highest Income Class	3.34e-10	0.00278



Coefficients	Normal	Transformed
Intercept	30.8	0.499
Fibre	-0.0055	-3.38e-07
Sugar	0.00491	1.59e-07
Sat Fat	-0.0225	-5.14e-07
Polyunsat Fat	-0.0531	-1.29e-06
Monounsat Fat	0.0403	8.94e-07
Fruit	-0.248	-6.11e-06
Veg	-0.00993	-1.29e-07

Coefficients	Normal	Transformed
Intercept	30.4	0.499
Fibre	-0.0378	-9.04e-07
Sugar	-0.0195	-4.98e-07
Sat Fat	0.0503	6.66e-07
Polyunsat Fat	-0.124	-3.18e-06
Monounsat Fat	0.0944	2.77e-06
Fruit	0.703	1.58e-05
Veg	0.172	4.13e-06

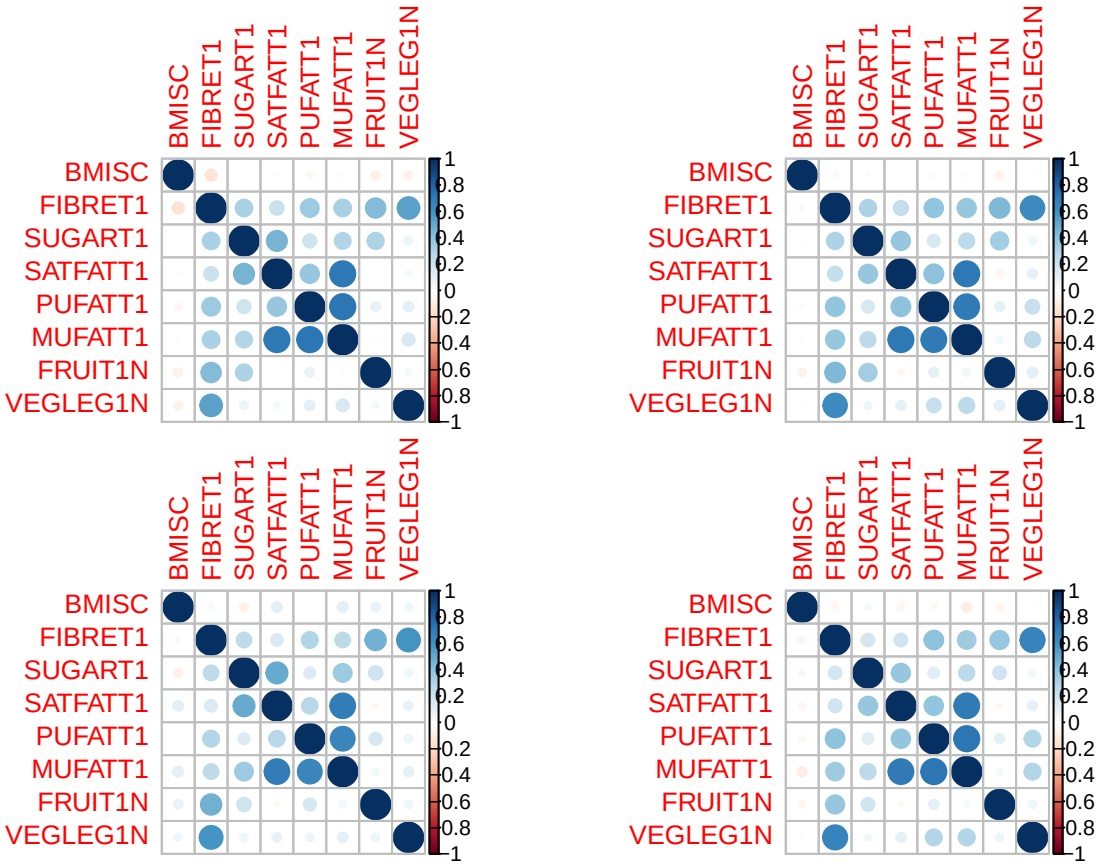
Unhealthy females in the lowest income class are characterised by their lower fibre consumptions and increased sugar consumption. They are more likely to fail to meet the recommended daily fruit and vegetable intake criteria.

Females belonging to the middle and higest income classes share some of the same characteristics with women in the lowest income class. These are manifested in the form of fibre levels, which are lower in all income

Coefficients	Normal	Transformed
Intercept	31.8	0.499
Fibre	-0.0382	-1.28e-06
Sugar	0.00778	2.44e-07
Sat Fat	0.0151	7.08e-07
Polyunsat Fat	0.0961	3.41e-06
Monounsat Fat	-0.0917	-2.86e-06
Fruit	-0.154	-6.31e-06
Veg	0.171	6.28e-06

classes as one becomes heavier and consequentially more unhealthy. Furthermore, fruit intake drops matched with a higher sugar intake. Some causes may include the recent surge of fruit juice as a replacement for whole fruit. These juices are mostly reconstituted and thus lack the necessary fibre required in a healthy diet. We will now attempt to portray some correlation plots for the variables of interest in this research section.

```
## corrplot 0.84 loaded
```



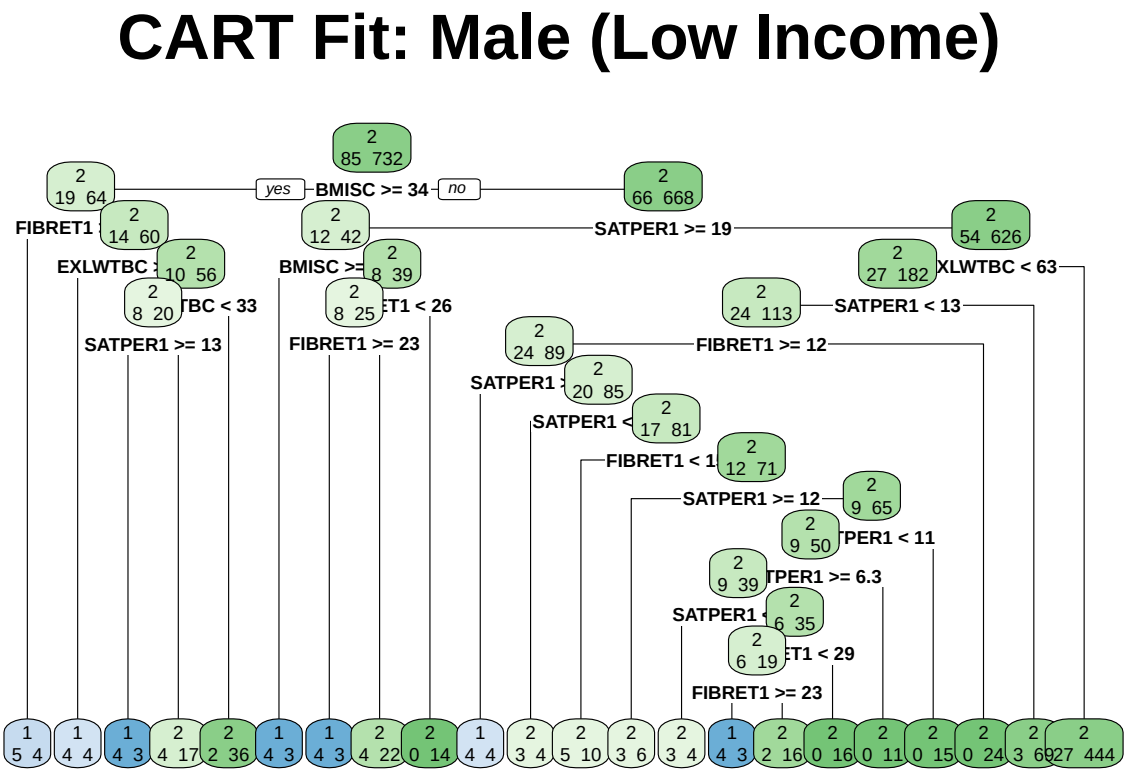
6 Fibre and other factors in ensuring good health

6.1 Question Focus:

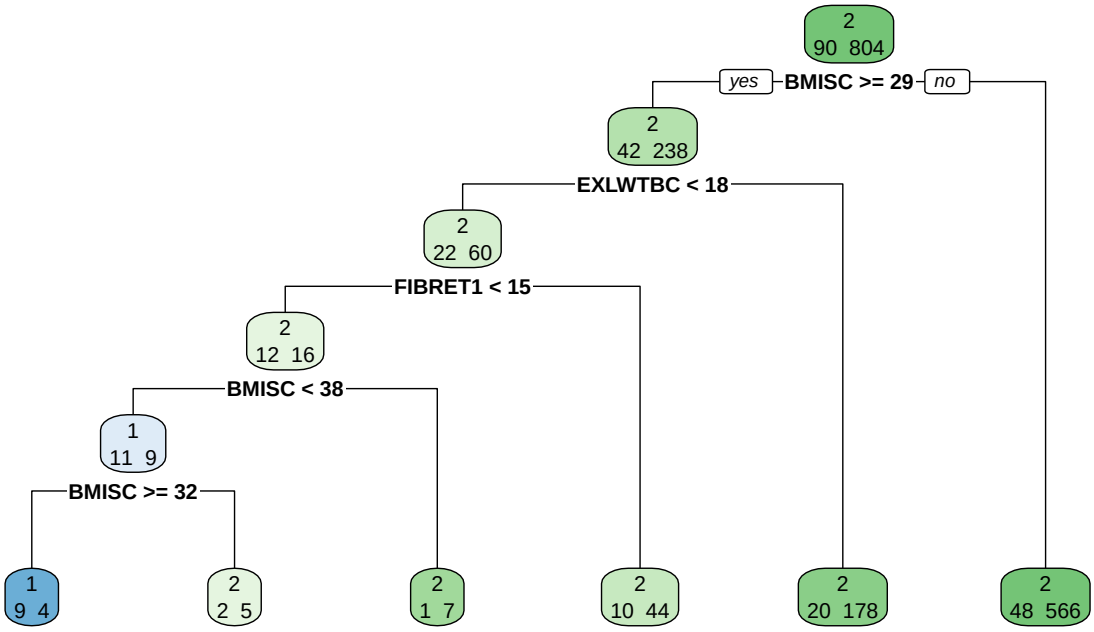
This final section attempts to investigate whether increasing fibre intake decreases the chance for an individual to develop diseases such as hypertension, high cholesterol levels or high blood sugar levels. We make links to the core question by paying particular attention to the trends of fibre across the 3 income levels. We have

chosen CART as the classifier for this section and will attempt to classify whether an individual suffers from the aforementioned diseases. Following the same reasons in the previous question, we omit the analysis for Males in Middle and Upper Income Levels due to the lack of observations which decreases statistical power.

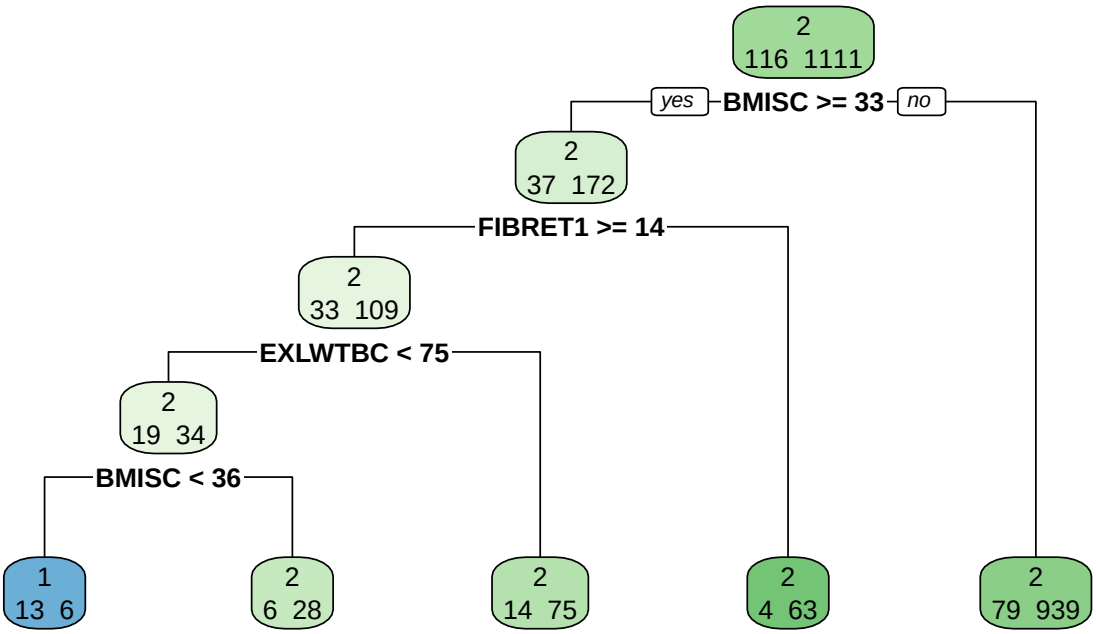
6.1.1 High Cholesterol



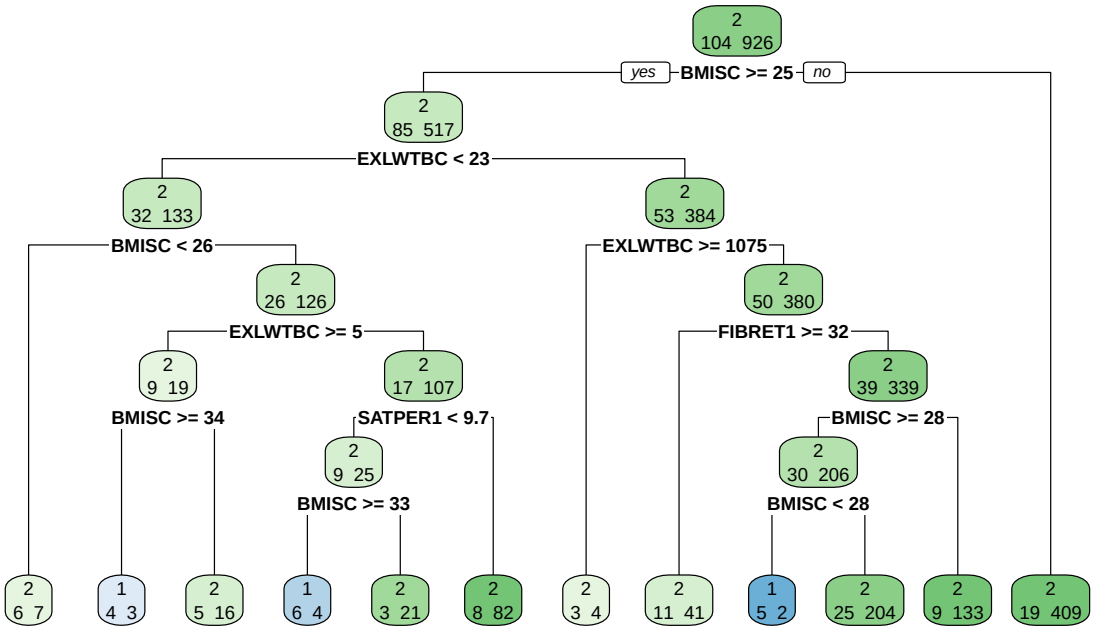
CART Fit: Female (Low Income)



CART Fit: Female (Average Income)



CART Fit: Female (High Income)



6.1.1.1 Meaningful Conclusions from High Cholesterol CART Analysis

The above classification has the following interpretation For males with low income, if the energy from saturated fat is less than 19, physical activity time is less than 63, saturated fat is greater than 13 and fibre intake is greater than 12 then you would have a low cholesterol level(24 cases), otherwise not.

For males with average income level ,if BMI level less than 27,physical activity time less than 5,energy from saturated fat is greater than 8 ,fibre lever is greater than 29, then you would have a low cholesterol level(20 cases), but under this condition if fibre level gets lower than 20, then you get a high cholesterol level. So we may suggest fibre plays a role for cholesterol in middle income level.

For Males with high income, if BMI level is less than 29 and greater than 22, fibre intake is greater than 11, then you get a low cholesterol level(43 cases). On the other hand, if BMI level is greater than 29, fibre intake greater than 19 and greater than 34, then you get a low cholesterol level(54 cases). We may conclude in the high income level, higher fibre intake relates to lower cholesterol levels.

For Females with low income, BMI level greater than 29, physical activity less than 18 and fibre intake greater than 15, then you get a low cholesterol level(44 cases).

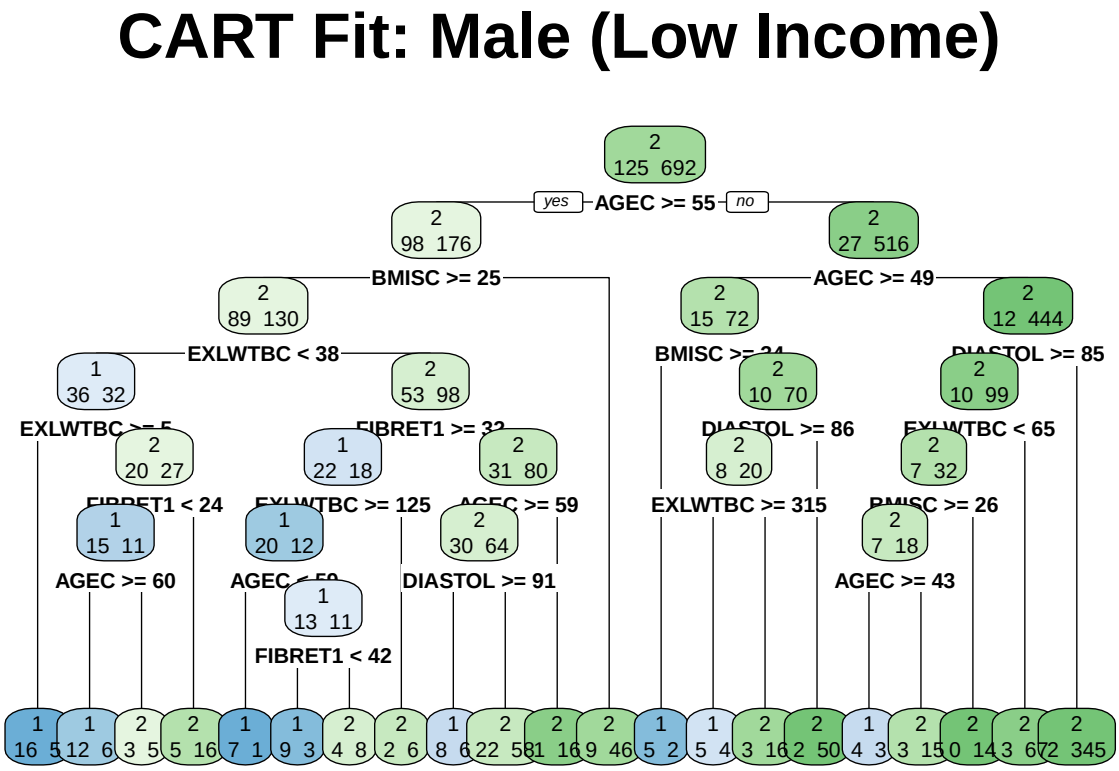
For Females with average income, BMI level greater than 33 and fibre intake greater than 14, then you get a low cholesterol level(63 cases). This seems similar to the result we get from low income level.

For Females with high income, BMI level greater than 25, physical activity greater than 23 and fibre intake greater than 32, then you get a low cholesterol level(41 cases). We still get similar results, so higher fibre intake gives lower cholesterol levels in Females.

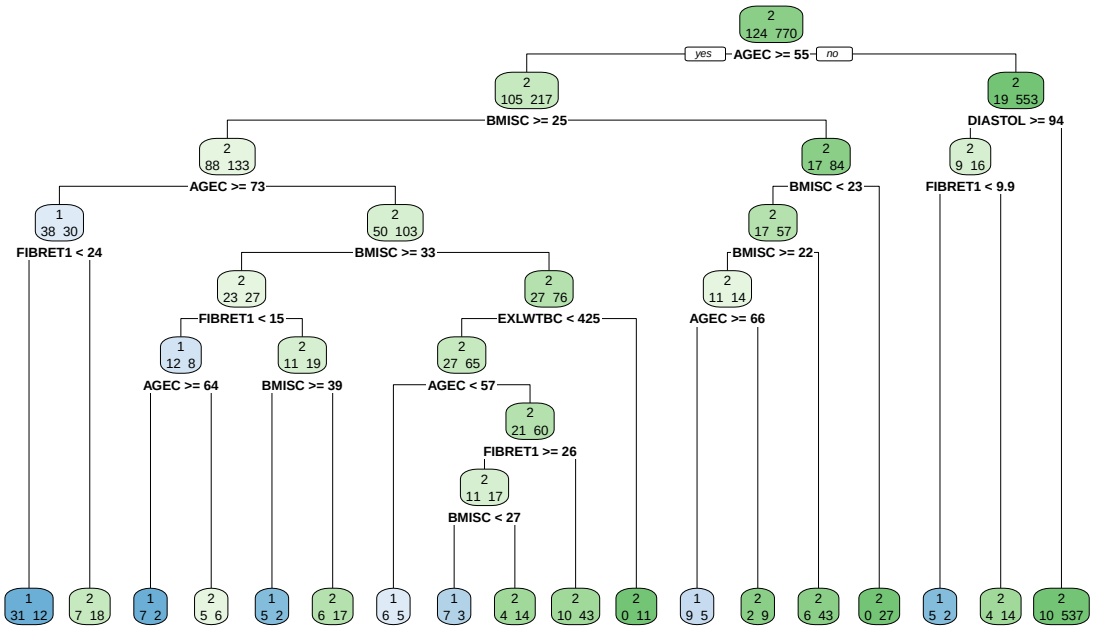
6.1.2 High Blood Sugar Levels

The classification tree for blood sugar levels possesses very low CV errors. This makes sense as from the data, there is an overwhelmingly negative response to testing positive for high blood sugar levels. Taking this into account, the classifier naturally selects option 2 ("No") all the time, resulting in the single node present for males and females across all three income classes.

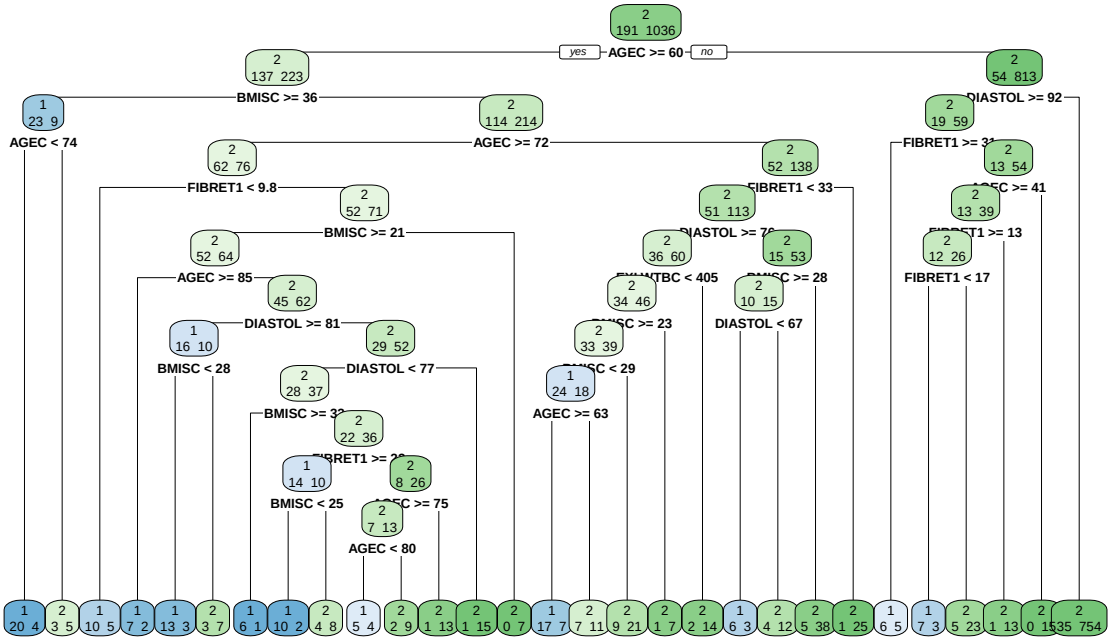
6.1.3 Hypertension



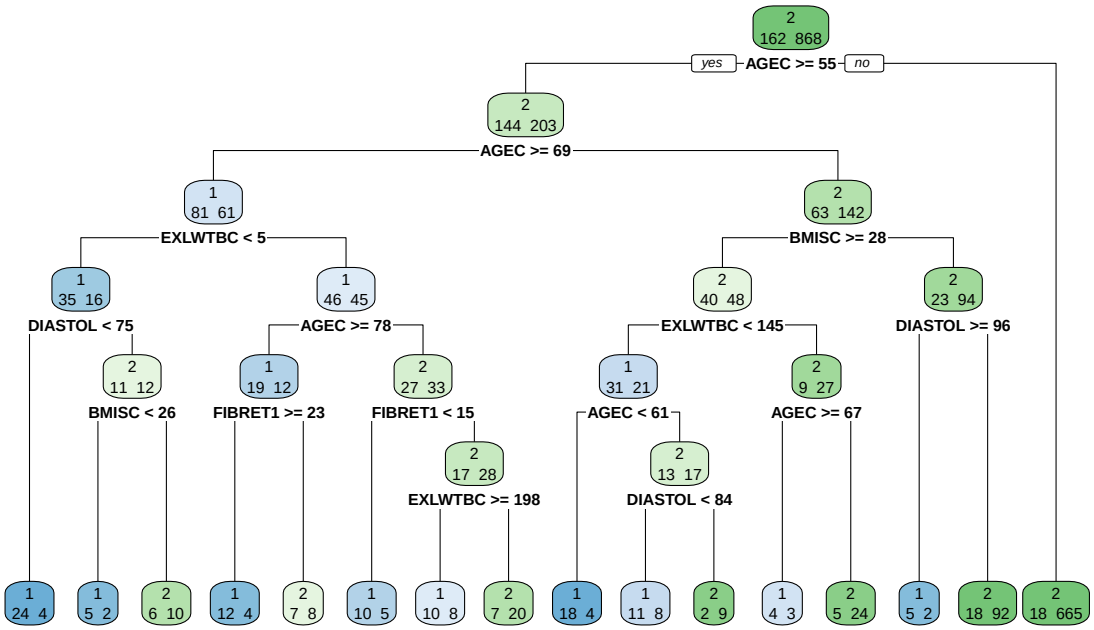
CART Fit: Female (Low Income)



CART Fit: Female (Average Income)



CART Fit: Female (High Income)



6.1.3.1 Meaningful Conclusions from Hypertension CART Analysis

Males with low income, an age greater than 55, a BMI greater than 25, physical activity greater than 38 minutes a week (but less than 125), fibre intake greater than 32 grams a day corresponds to a low chance of hypertension (6 cases).

Males with high income, an age greater than 59, a BMI greater than 31 and a fibre intake less than 13 grams a day corresponds to a high chance of hypertension (15 cases).

Females with low income, an age less than 55, diastolic blood pressure greater than 94 mm/Hg and a fibre intake greater than 24 grams a day are less likely to suffer from hypertension (18 cases).

Females with an average income, an age less than 60, diastolic blood pressure greater than 92 mm/Hg and a fibre intake greater than 31 grams/day are highly likely to suffer from hypertension (6 cases). So fibre intake may play a less important role here.

For Female with high income level,if age is less than 69, BMI level is less than 28 and diastolic blood pressure less than 96, then get a high hypertensive disease chance(92 cases). Similarly fibre intake may play a less important role here.

7 Conclusion

Within this report, we have investigated the link between an individual's income and diet. With the report agenda being to raise awareness as to the health benefits of incorporating high levels of dietary fibre we have incorporated the use of statistical techniques in order to deduce the answers to the questions that we proposed. We have also built classifiers, borrowing feature variables that literature dictates as important

in the prediction of whether an individual would suffer from high cholesterol levels. We then underpinned the negative trends of fibre within unhealthy inviduals and found that this trend was consistent regardless of income category. As such a final CART classification algorithm was implemented in order to deduce the sufficient amount of fibre that an individual should intake daily in order to maintain their health.

References

Aller, de Luis, R. 2004. "Effect of Soluble Fiber Intake in Lipid and Glucose Leves in Healthy Subjects: A Randomized Clinical Trial." *Diabetes Research and Clinical Practice* 65 (1): 7–11.

Buttriss, & Stokes, J. L. 2008. "Dietary Fibre and Health: An Overview." *Nutrition Bulletin* 33 (3): 186–200.

Delzenne, & Cani, N. M. 2005. "A Place for Dietary Fibre in the Management of the Metabolic Syndrome." *Current Opinion in Clinical Nutrition & Metabolic Care* 8 (6): 636–40.

Foundation, Australian Health. 2017. "Overweight and Obesity Statistics." https://www.heartfoundation.org.au/about-us/what-we-do/heart-disease-in-australia/overweight-and-obesity-statistics.

Paris Michel-Ange), CNRS (Délégation. 2014. "How Fiber Prevents Diabetes, Obesity." https://www.sciencedaily.com/releases/2014/01/140114090822.htm.

ResearchGate. 2016. "How Large Should a Sample Size Be for Multiple Regression? What Is the Minimum?" https://www.researchgate.net/post/How_large_should_a_sample_size_be_for_multiple_regression_What_is_the_minimum.

Slavin, & Green, J. 2007. "Dietary Fibre and Satiety." *Nutrition Bulletin* 32: 32–42.

Van de Vijver, Van den Bosch, L.P.L., and R.A. Goldbohm. 2009. "Whole-Grain Consumption, Dietary Fibre Intake and Body Mass Index in the Netherlands Cohort Study." *European Journal of Clinical Nutrition* 63 (1): 31.