STAT3014/STAT3914 Major Project Report

Dong Luo, Steph Standfort, and Anqi Chi

26 October 2018

Contents

1 Executive summary

Obesity affects approximately one third of the Australian populated aged 18 years and over (Australian Bureau of Statistics, 2015), posing major health risks to individuals. Obesity has been demonstrated to be related equally to genetics and environmental factors including diet. A person who is dieting has been defined as one who consumes macronutrient groups in an amount lying outside of the acceptable macronutrient distribution ranges (National Health and Medical Research Council, 2017). Here we are interested in whether consuming a diet outside of normal macro-nutrient ranges may be related to Body Mass Index BMI, and thus obesity, and whether a person's demographics, for example sex, age, waist size, and socioeconomic status is predictive of a person's choice of diet. To investigate this, we developed three research questions.

1.1 Research Questions

1. What are the most common diet types among adults in the sample, categorised by the macro- nutrients fat, proteins and carbohydrates.
2. To see if we could predict obesity measured using BMI based on eight variables, including BMR, energy intake, sex, and time spend sedentary.
3. Investigated whether people on different diets have different characteristics and demographics, for example socio- economic status, age, sex, Basal Metabolic Rate (BMR), and BMI.

Throughout the analysis, we focussed on the macro- nutrients protein, fat and carbohydrates. These were because three common types of diets trends have been identified (Kossoff, Turner, Doerrer, Cervenka, Henry, 2016) which are all combinations of different levels of each protein, fat, and carbohydrates. These diets are the Keto diet (high fat, low carbohydrates, medium protein), the Atkins diet (high protein, low carbohydrates, medium fat), and the Low Fat-High Carb diet.

1.2 Main Findings

- The estimated proportion of people on Atkins, Keto and Low Fat-High Carb diet are 7.1%, 17.8%, and 1.7% repectively.

- The following variables were found to increase the probability of being obese: being female. being on a diet, higher BMR, lower energy intake, less total minutes spent sitting or lying down, having a low percentage of energy from carbohydrate.
- Some of the findings were unexpected compared to previous literature. For example, men tended to have higher carbohydrate diets compared to women, and high protein diets had a higher proportion of people from low SES deciles compared to normal SES deciles. These findings are further explored in section 3 of this report.

2 Methodologies

2.1 Data Cleaning

We used the code provided by John Ormerod for the STAT3014 Lab 2. Missing values became NAs. BMI, gender, and SES were recoded as binary. Table 1 contains details of the variables used in this report.

Table 1: Description of Variables

	Class	Description
BMI	Binary	0 = BMI < 30, 1 = BMI > 30
BMR	Numeric	Kilojules burnt per day
Waist	Numeric	Waist size in centimeters
Sedentary	Numeric	Minutes spent sitting or lying down
Sex	Binary	0 = female, 1 = male
SES	Binary	Socio- economic status where 0 = lower SES, 1 = normal SES
Age	Numeric	Age in years
Energy intake	Numeric	Energy (kilojules) per day
Diet status	Binary	0 = not on a diet, 1 = on a diet
Carbohydrates	Numeric	Percentage of energy from carbohydrates
Protein	Numeric	Percentage of energy from Protein
Fat	Numeric	Percentage of energy from fat

2.2 Proportion of the diets

We excluded the observations where age was less than 18. This was because our research questions focussed on obesity which we determined using the BMI. However BMI does not apply to children, so including under 18s may lead to misleading results.

We cut each of the continuous variables `CHOPER1`, `FATPER1`, and `PROPER1`, which are the percentages of total energy coming from carbohydrates, fat,and protein into three distinct levels,low, medium, and high. Table 2 shows the percentages of total energy used to divide the groups.

Table 2: Division of Macro- nutrients by Total Energy Percentage

	Carbohydrates	Fat	Protein
Low (%)	[0,45]	[0,20]	[0,15]
Medium (%)	(45,65]	(20,35]	(15,25]
High (%)	(65,100]	(35,100]	(25,100]

We divided the data this way because we were interested in the mean proportion of each diet, so that we could identify the most popular diets within our sample.

2.3 Research Question 1 : Popular Diet Types

Table 3: Proportion of the top 6 diet types

proportion	fat	carb	protein
0.1902491	medium	medium	medium
0.1785904	high	low	medium
0.1383148	medium	low	medium
0.1267621	medium	medium	low
0.0714361	medium	low	high
0.0557499	high	low	low

Table 3 shows the sample prorpotions of the six most popular diet types.
We fitted log-linear models to see if there was any independence underlying the three-way contingency table.

2.3.1 Structural Zeroes

We noticed that there were some cells with value zero in our table. We treated the zeroes as structural zeroes (impossible combinations) since the variables `CHOPER1` (carbohydrate), `FATPER1` (fat), and `PROPER1` (protein) in the original dataset stand for the proportions. This means some of the diets listed in the table, such as high carbohydrate, high fat,high protein were not possible as the sum of the three proportions would exceed 100. We removed the cells of structural zeroes from the and model the incomplete table.

2.3.2 Log-Linear Models

There were 20 out of 27 cells with positive entries. The null model will have $20 - 1 = 19$ degrees of freedom and the additive model will have 13 degrees of freedom.

We started with the additvie model, which stands for the complete independence.

We found the deviance for the additive model (including all three factors) was 3472.4957366. We compared it with the models with one two way interaction. There were three such models, and their residual deviance is shown as in Table 4.

The model with `carb:fat` interaction had the lowest deviance. The difference of deviance between this model and the additive model was 1827.5168456. The two models are nested and when we compared them, the H_0 was the additive model (the smaller model).

$$\text{M1}\,(H_0) : \hat{\mu} = \beta_0 + \alpha_i + \beta_j + \gamma_k,$$

where α_i, β_j, and γ_k denote the i^{th}, j^{th} and k^{th} group of carbohydrate, fat, and protein levels. The alternative hypothesis is

$$\text{M2}\,(H_A) : \hat{\mu} = \beta_0 + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij}$$

Under H_0, the difference in deviance followed a χ^2 distribution whose degrees of freedom equals the difference in residual degrees of freedom of the two models.

If our table was complete, we would expect the difference in degrees of freedom to be $(3 - 1) \times (3 - 1) = 4$, since each factor has 3 levels.

Table 4: Deviances of models with one interaction term

	CF	CP	FP
Residual Deviance	1644.979	2828.39	3356.749

Here the difference in degrees of freedom is 3. This is because there was a structural zero in the marginal table of carbohydrate and fat (we cannot have high carbohydrate and high fat at the same time), so we were only adding three parameters when fitting the model. A coefficient of an interaction level in the `glm` function output will be `NA`.

The *p*-value for the test is close to 0, so we would reject the null hypothesis and prefer the model with `carb:fat` interaction. This model with one interaction term stands for the block independence.

We use the same procedure to test the models with more interaction terms, and the result of the test is summarised in Table 5.

Table 5: Deviances for Poisson log-linear models

Type	Model	Deviance	d.f.
Completely Independence Model	C+F+P	3472	13
Block Independence Model	P+CF	1645	10
	F+CP	2828	10
	C+FP	3357	9
Conditional Independence Model	CF+CP	668	7
	CF+FP	1566	6
	CP+FP	2752	6
Uniform Association Model	CF+CP+FP	449	3
Saturated Model	CFP	0	0

The letters C, F, and P in the table represent the variables carbohydrate, fat, and protein respectively. Symbols such as FP indicate we are including the interactin term `fat:protein`, and when we include an interaction term we must also include the variables used to compute the interaction. Similarly, when we include the three-way interaction term CFP (`carb:fat:protein`), we include all two-way interactions as well as all of the three variables.
The deviance test suggests that we shall choose the saturated model, which implies

$$\hat{\mu}_{ijk} = y_{ijk}.$$

Therefore, the proportion for each diet can be estimated by the sample proportions.

2.4 Research Question 2 : Predicting Obesity

To predict the obesity (level of BMI is greater than 30), we chose eight variables as the predictors, which include BMR, energy intake, total minutes spent sitting or lying down, sex, whether currently on a diet, percentage of energy from carbohydrate, percentage of energy from total fat and percentage of energy from protein.

We applied linear discriminant analysis, classification and regression tree and logistic regression to investigate our question and estimated using 10-fold cross-validation.

We assign the level of BMI greater than 30 as number 1 which represent obesity and the remaining as number 0 to represent normal people. we also assign people currently on a diet as number 1 and those not a on a diet as number 0, and ignore the situation not known if currently on a diet and not applicable. Thus, we treat all variables that we are interested as numerical variables so that we increase the effectiveness of our analysis and the interpretability of our results.

2.4.1 LDA

For our prediction attempt, we first use linear discriminant analysis rather than quadratic discriminant analysis because the Interpretation for QDA is more difficult, but the group means suggest identical interpretations as for LDA.

After fitting the LDA method, the output (Table 6) indicates the following interpretations for each of the variables. People who have higher BMR are more likely to be obese as the coefficient of linear discriminant 0.001249 is positive. Lower energy intake increases obesity probability since the coefficient of linear discriminant -0.434204 is negative. People who have lower total minutes spent sedentary are prone to be obese. Females are more likely to have obesity and people on a diet are more likely to be obese compared with people not on a diet. People who have high fat or high protein diet type are more likely to be obese while the probability of obesity reduces with high carbohydrate diet type.

Table 6: Coefficients of Predictors in LDA

	Estimated Coefficients
BMR	0.001249
Energy Intake	-0.434204
Total mins spent sedentary	-0.000046
Sex	-2.093857
Whether currently on a diet	0.363241
Carbohydrate diet	-0.005032
Fat diet	0.001558
Protein diet	0.002124

2.4.2 CART

In addition, we proceeded to fit a CART for obesity by the dependent variables we selected before and then we got the output below of CART Fit for Obesity. The tree could be explained precisely by Table 7 and Table 8. We found that there is no difference of obesity rate between male and female whose level of BMR are lower than 6103. The obesity rate is 68.5 percent for male with the level of BMR greater than 8297 while 80.1 percent of female with BMR greater than 6103 are obese. 83.3 percent of male whose BMR are between 6103 and 8297 are normal. 54.2 percent of female whose BMR are between 6103 and 6390 are normal. Based on this analysis, we may conclude that at the same level of BMR, female are more likely to have obesity than male.

Table 7: Obesity Rate for Male with BMR greater than 6103

	6103 < BMR < 8297	BMR > 8297
Estimated obesity rate	6.7%	68.5%

Table 8: Obesity Rate for Female with BMR greater than 6103

	6103 < BMR < 6390	BMR > 6390
Estimated obesity rate	45.8%	80.1%

2.4.3 Logistic regression

As we are interested in a special case on a generalised linear model for binary data which is obesity in our case, we fit a logistic regression model on the same data. The full model has the coefficients for BMR, energy intake, total minutes spent sedentary, sex, whether currently on a diet, percentage of energy from

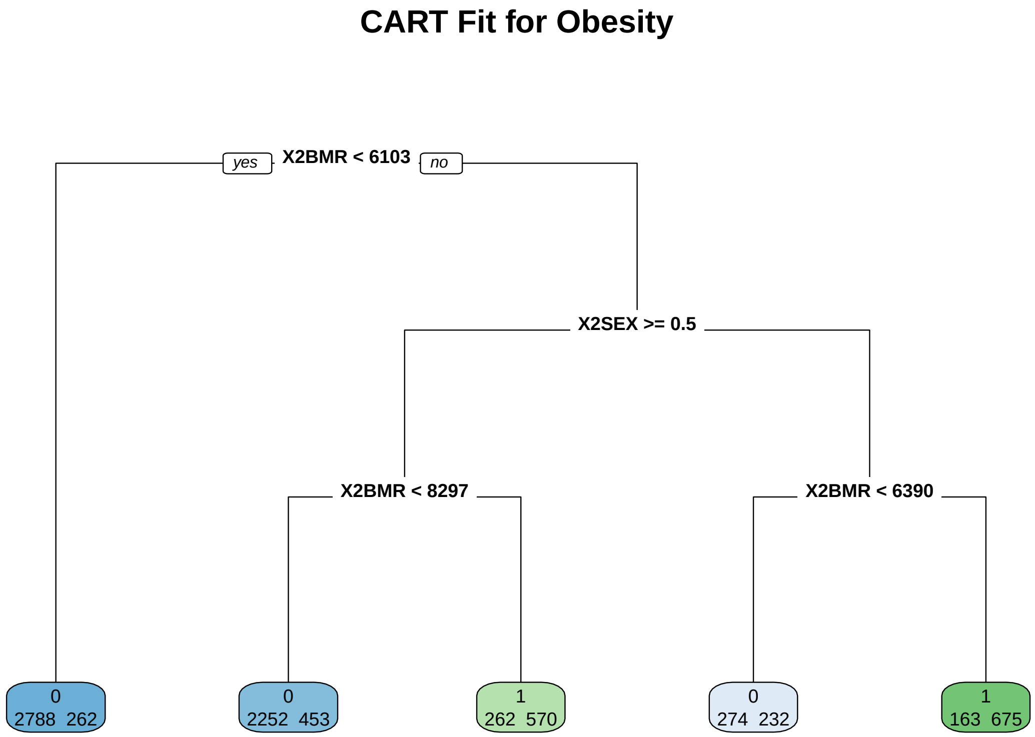

Figure 1: CART Fit for Obesity

carbohydrate as statistically significantly different from zero at the 0.05 level. Table 9 shows the estimated coefficients of the logistic regression.

Table 9: Estimated Coefficients of the Logistic Regression

	Variable	Estimated Coefficient
(Intercept)	1	-10.925622
BMR	x_1	0.001879
Energy Intake	x_2	-0.711322
Total mins spent sedentary	x_3	-6e-05
Sex	x_4	-3.422641
Whether currently on a diet	x_5	0.380154
Carbohydrate diet	x_6	-0.010135
Fat diet	x_7	0.001507
Protein diet	x_8	-0.000998

The fitted model is

$$\begin{aligned}\mathrm{logit}(p) = &- 17.770903 + 0.001879x_1 - 0.711322x_2 - 0.00006x_3 - 3.422641x4 \\ &+ 0.380154x_5 - 0.010135x_6 + 0.001507x_7 - 0.000998x_8\end{aligned}$$

2.4.4 CV Errors

Table 10: Summary of the CV Errors

Methods	Errors
LDA	18.232
CART	17.551
Logistic Regression	17.778

Finally, we compare these three methods by computing repeated 10-fold cross-validation errors. Table 10 shows that the CV error for CART is 17.551 percent which is better than LDA and Logistic Regression. The effect of different methods on obesity are relatively consistent.

2.5 Research Question 3 : Comparing Demographics Between Diets

2.5.1 Model Selection

This section worked out the best diets to model different dependent variables; BMI, BMR, waist size (cm), mintues spent sedentary, sex, SES, and age. The focus of this area was to fit eight key demographics using protein, fat, and carbohydrate diets. Each of the variables was chosen through discussion with NUTM3001 students, and were modelled using linear regression for numerical variables and logistic regression for binary variables.

Prior to beginning model selection, assumptions for normality of errors and homogeneity of variance were assessed using Q-Q plots and box plots of residuals, and a residual versus fitted plot was used visually assess homogeneity of variance and independence. An example of the graphs used has been provided in Figure 2 showing the diagnostic plots for the dependent variable BMR, excluding the boxplot of residuals. For variables BMR and sedentary the normality assumption was not met. We decided to apply a log transformation to BMR and a square root transformation to minutes sedentary to satisfy normality.

Once the the assumptions were met, we began model selection. We used residual deviance tests and AIC Stepwise model selection to find the “best” model for continuous dependent variables and the F test and AIC Stepwise model selection for binary dependent variables. For five of the dependent variables the “best” model

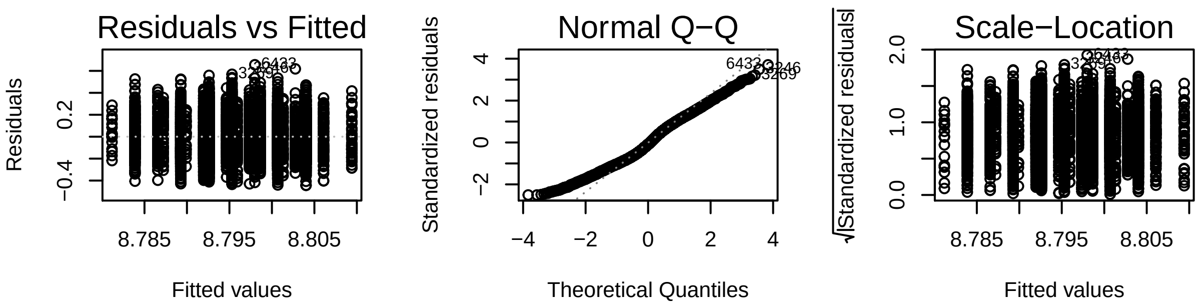

Figure 2: Model Diagnostics

was found to be the same using AIC and the residual deviance or F test. For BMR, SES, and minutes spent sedentary the models did not match. When the models did not match, the model selected by the F test was used as the "best" model because the AIC has been claimed to typically prefer overly complex models.

For each dependent variable using the F or deviance tests for nested models, we tested the null hypothesis that the variable was best modelled by a smaller model beginning with the null model, compared to the alternative hypothesis that the variable was best explained by a model with a single extra term. Linear and logistic regression tested the null hypothesis that adding an extra term will not give a better fit of the variable, compared to the alternative hypothesis that adding an extra term will increase the goodness of fit on the model.

A 5% significance level was chosen as a threshold for the inclusion of the model variables. The resulting "best" models explaining each of the dependent variables is shown in Table 11.

Table 11: Best models from model selection

	Deviance	AIC	F- test
BMI	P	P	
BMR		C+PF	P
Waist (cm)		P+C	P+C
Mins sedentary		P+F+C	P+F+C
Sex (proportion of male)	C+PF	C+PF	
SES (proportion of high SES)	C	C+PF	
Age		P+F+C	P+F+C

2.5.2 Demographics for diets

Table 12 presents the key characteristics of each of the three diets investigated. For BMR, waist measurement, minutes spent sedentary, and age, the mean of each group was reported. For BMI, sex, and SES respectively, the probability of a person being overweight, male, or in the normal decile Index of Relative Socio-Economic Disadvantage has been reported.

2.5.3 A couple of key things to note:

- BMI: We found the proportion of people who were obese was highest in the high protein diet. A recent CSIRO Protein Balance Report suggests that a high protein diet can improve body composition (Noakes, 2018). Whilst this is the opposite of our findings, it is possible that individuals with a larger waist circumference may have been consuming a high protein diet with the prospective goal of losing

Table 12: Demographics for 8 key variables

	Atkins	Keto	Low Fat, High Carb
BMI	0.358	0.282	0.282
BMR	6724.316	6713.304	6897.837
Waist (cm)	94.510	93.264	94.489
Mins sedentary	2259.420	2388.410	2054.230
Sex	0.416	0.384	0.375
SES	0.237	0.286	0.197
Age	50.497	48.333	44.031

weight. It is unknown whether the diet has been successful in this regard as the AHS only records one point in time. Another consideration may be that people on high protein diets could be focussed on muscle development which could result in a higher body weight, and consequently a higher BMI. More data would need to be collected around subjects exercise activites and purpose for diet choices to test these theories.

- Waist: People on high protein diets had the largest measn waist measurement (95.204cam) compared to any of the other diets test, including Atkins, Keto, and the low carbohydrate/ high fat diet. A persons waist size can show whether a person is carrying excess fat around their middle and can be an indicator of the level of internal fat deposits covering internal organs. In conjuction with BMI, they can help determine a persons risk of health issues, such as stroke or heart disease.
- Sedentary: The high fat low carbohyrate diet (keto diet) had the highest mean sedentary minutes of 2388.410, which equates to 39.8 hours of being sedentary over two days. This diet is often advertised as an effective diet for athletes, however this sample was not targeted towards athletes. It may have been interesting to look further into this comparing people who said that they were on a diet compared to not on a diet who fell into the Keto diet, and to look into their respective minutes spent sedentary and minutes spent exercising.
- Sex: Within our dataset 46% of subjects were male yet we found that the probability of being male on a high carbohyrate, medium fat, and medium protein diet was 60%.
- SES: Despite having a lower proportion of normal SES participants in the study, there was a high proportion who had a high carbohyrate diet but normal fat and protein. This is a surprising result as much literature supports low SES rather than normal and high SES having high carb diets since carbohyrates are cheaper than protein, and might be consumed ecessively in order to to meet dietary protein needs (Brooks, Simpson & Raubenheimer, 2010). However some literature comments on the inconsistenices of results related to protein intake in the context of lower SES (Darmon,N., Drewnowski,A.,2008).

2.6 Limitations

We identified several limitations in our project. Stated below:

- This project only focussed on 11 out of the available 144 variables provided in the dataset. Had we had more time, these other variables could have been further investigated to make more accurate or interesting findings.
- The assumption of Poisson distribution when fitting the log-linear model might be wrong. Since, the number of total observations is fixed, the true distribution is multinomial.
- The continuous variables BMI and socio- economic status (SES) were manually split into two groups (normal/ overweight and obese, and low SES and normal SES). This could result a loss of information, or less accurate interpretation of the data.

- Some of the assumptions in model selection using linear regression were not met so log and square root transformations were applied in some cases to the data. This could affect the interpretation of the data.
- Stepwise tests were used for model selection where the test statistics may not follow the F or chi-squared distributions (Flom, P.L, Cassell, D.L, 2007)
- An automated stepwise AIC was used in model selection. Although it was used in conjunction with deviance tests and the F test, there is some criticism that the AIC only selects a locally optimal model (Thayer, J.D, 1990).
- The three macro- nutrients being looked at are comprised of specific subcategories, for example fat is made up of different types of fats including trans- fats (which are bad for you) or polysaturated fats (which are good for you) and these have not been investigated here, but the breakdown into subcategories may lead to further explanation of results.
- We only provided point estimation of the fitted proportions in logistic regression rather than a confidence interval.

References

Brooks, Simpson, R.C., and D. Raubenheimer. 2010. "The Price of Protein: Combining Evolutionary and Economic Analysis to Understand Excessive Energy Consumption." *Obesity Reviews* 11 (12): 887–94.

Cassell, Flom, D.L. 2007. "Stopping Stepwise: Why Stepwise and Similar Selection Methods Are Bad, and What You Should Use."

Darmon, A., N. & Drewnowski. 2008. "Does Social Class Predict Diet Quality?" *The American Journal of Clinical Nutrition* 87 (5).

Kossoff, Turner, E.H. 2016. "The Ketogenic and Modified Atkins Diet: Treatments for Epilepsy and Other Disorders." *Obesity Reviews* 11 (12). Springer Publishing Company: 887–94.

Noakes, M. n.d. "Protein Balance: New Concepts for Protein in Weight Management." CSIRO, Australia.

Thayer, J.D. 1990. "Implementing Variable Selection Techniques in Regression." *Annual Meeting of the American Educational Research Association*, 16–20.