Clara

Anqi Chi SID:460204008

22/10/2018

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.

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 1) 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 1: 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

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 2 and Table 3. 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.

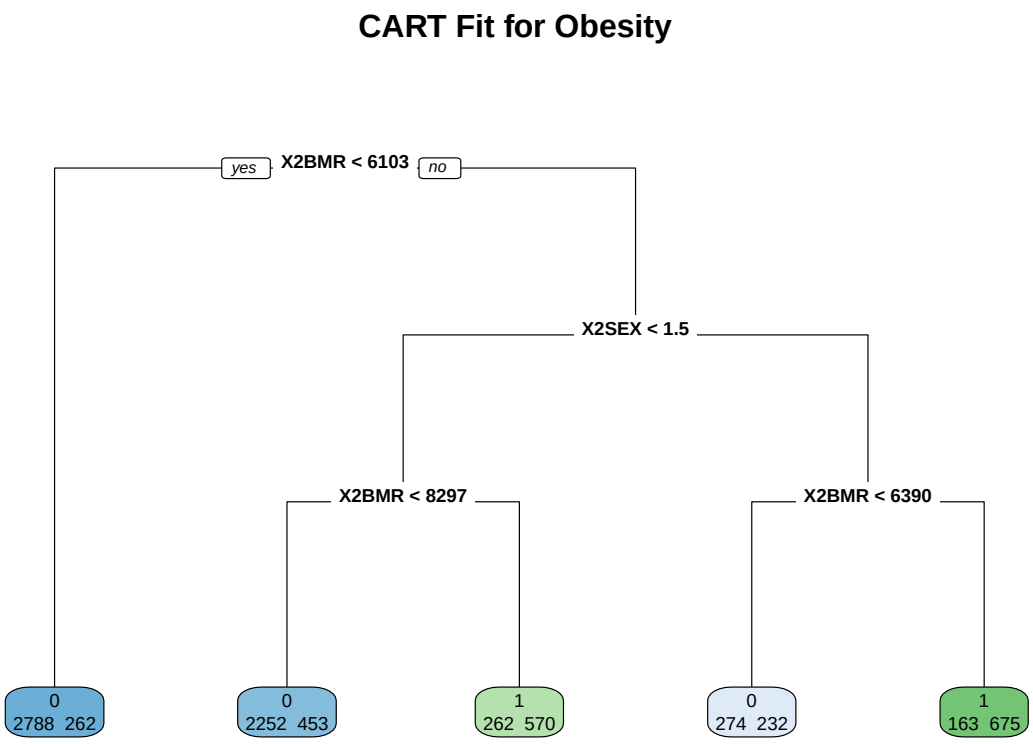

Figure 1: CART Fit for Obesity

Table 2: Obesity Rate for Male with BMR greater than 6103

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

Table 3: Obesity Rate for Female with BMR greater than 6103

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

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 carbohydrate as statistically significantly different from zero at the 0.05 level. Table 4 shows the Estimated Coefficients of the Logistic Regression.

Table 4: Estimated Coefficients of the Logistic Regression

	Variable	Estimated Coefficient
(Intercept)	1	-17.770903
BMR	x_1	0.001879
EIBMR1	x_2	-0.711322
ADTOTSE	x_3	-6e-05
SEX	x_4	3.422641
BDYMSQ04	x_5	0.380154
CHOPER1	x_6	-0.010135
FATPER1	x_7	0.001507
PROPER1	x_8	-0.000998

The fitted model is

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

where p represent the probability of obesity.

where $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8$ represents BMR, EIBMR1, ADTOTSE, SEX, BDYMSQ04, CHOPER1, FATPER1, PROPER1 respectively.

From this fitted model, we noticed that the effects of the significant variables on obesity are the same with LDA, and we can get similar explanations with LDA method but differs that high percentage of energy from protein are less likely to be obese.

summary

Table 5: 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 5 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.