Clara

Anqi Chi SID:460204008

22/10/2018

Data Clean

```
## [1] 12153    145

##     BDYMSQ04           BMR             SEX           EIBMR1
##  Min.   :0.0000   Min.   : 4275   Min.   :1.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.: 5734   1st Qu.:1.000   1st Qu.:0.9381
##  Median :0.0000   Median : 6504   Median :2.000   Median :1.2371
##  Mean   :0.1382   Mean   : 6710   Mean   :1.524   Mean   :1.3064
##  3rd Qu.:0.0000   3rd Qu.: 7595   3rd Qu.:2.000   3rd Qu.:1.6033
##  Max.   :1.0000   Max.   :12759   Max.   :2.000   Max.   :5.8661
##     ADTOTSE          CHOPER1          FATPER1          PROPER1
##  Min.   :   0    Min.   : 0.00   Min.   : 0.00   Min.   : 0.00
##  1st Qu.:1410    1st Qu.:36.11   1st Qu.:24.99   1st Qu.:14.14
##  Median :2165    Median :43.51   Median :30.88   Median :17.40
##  Mean   :2358    Mean   :43.22   Mean   :30.83   Mean   :18.34
##  3rd Qu.:3180    3rd Qu.:50.48   3rd Qu.:36.53   3rd Qu.:21.38
##  Max.   :9180    Max.   :98.73   Max.   :82.87   Max.   :63.65
```

LDA

Fit the LDA method

```
##      BMR            EIBMR1          ADTOTSE          SEX
##  Min.   : 4275   Min.   :0.0000   Min.   :   0    Min.   :1.000
##  1st Qu.: 5734   1st Qu.:0.9381   1st Qu.:1410    1st Qu.:1.000
##  Median : 6504   Median :1.2371   Median :2165    Median :2.000
##  Mean   : 6710   Mean   :1.3064   Mean   :2358    Mean   :1.524
##  3rd Qu.: 7595   3rd Qu.:1.6033   3rd Qu.:3180    3rd Qu.:2.000
##  Max.   :12759   Max.   :5.8661   Max.   :9180    Max.   :2.000
##     BDYMSQ04          CHOPER1          FATPER1          PROPER1
##  Min.   :0.0000   Min.   : 0.00   Min.   : 0.00   Min.   : 0.00
##  1st Qu.:0.0000   1st Qu.:36.11   1st Qu.:24.99   1st Qu.:14.14
##  Median :0.0000   Median :43.51   Median :30.88   Median :17.40
##  Mean   :0.1382   Mean   :43.22   Mean   :30.83   Mean   :18.34
##  3rd Qu.:0.0000   3rd Qu.:50.48   3rd Qu.:36.53   3rd Qu.:21.38
##  Max.   :1.0000   Max.   :98.73   Max.   :82.87   Max.   :63.65
## Call:
## lda(y ~ ., data = X1, subset = 1:n)
##
## Prior probabilities of groups:
##         0         1
## 0.7236162 0.2763838
##
## Group means:
```

```
##         BMR    EIBMR1  ADTOTSE       SEX  BDYMSQ04  CHOPER1  FATPER1  PROPER1
## 0 6426.343 1.380460 2313.456 1.520822 0.1073358 43.38162 30.74535 18.05556
## 1 7452.157 1.112322 2475.082 1.533303 0.2189781 42.79542 31.06969 19.08146
##
## Coefficients of linear discriminants:
##                     LD1
## BMR        1.248732e-03
## EIBMR1    -4.342037e-01
## ADTOTSE   -4.590553e-05
## SEX        2.093857e+00
## BDYMSQ04   3.632406e-01
## CHOPER1   -5.031900e-03
## FATPER1    1.558290e-03
## PROPER1    2.123950e-03
```

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

Comment:

The above output suggests the following interpretations for each of the variables. * People who have higher BMR are more likely to obese * Lower EIBMR1(Energy intake) increases obese probability (since the coefficient -0.434 is negative). * Lower ADTOTSE(Total mins spent sitting or lying down) increases obese probability (since the coefficient -4.590553e-05 is negative). * Sex = 1 for males. So females are more likely to obese * People on a diet are more likely to obese compared with people not on a diet * People who have high fat and high protein diet type are more likely to obese * People who have high carbon diet type decrease the probability of obesity

CV error

```
#cross validation error for LDA.
res.lda = cv.da(X1,y,method="lda",V,seed=1)
res.lda
```

```
## [1] 0.1823225
```

Comment: The CV error for LDA is 18.232 percent

CART

```
## n= 7931
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 7931 2192 0 (0.72361619 0.27638381)
##    2) X2BMR< 6102.75 3050   262 0 (0.91409836 0.08590164) *
##    3) X2BMR>=6102.75 4881 1930 0 (0.60458922 0.39541078)
##      6) X2SEX< 1.5 3537 1023 0 (0.71077184 0.28922816)
##       12) X2BMR< 8297.25 2705   453 0 (0.83253235 0.16746765) *
##       13) X2BMR>=8297.25 832   262 1 (0.31490385 0.68509615) *
##      7) X2SEX>=1.5 1344   437 1 (0.32514881 0.67485119)
##       14) X2BMR< 6389.5 506   232 0 (0.54150198 0.45849802) *
##       15) X2BMR>=6389.5 838   163 1 (0.19451074 0.80548926) *
```
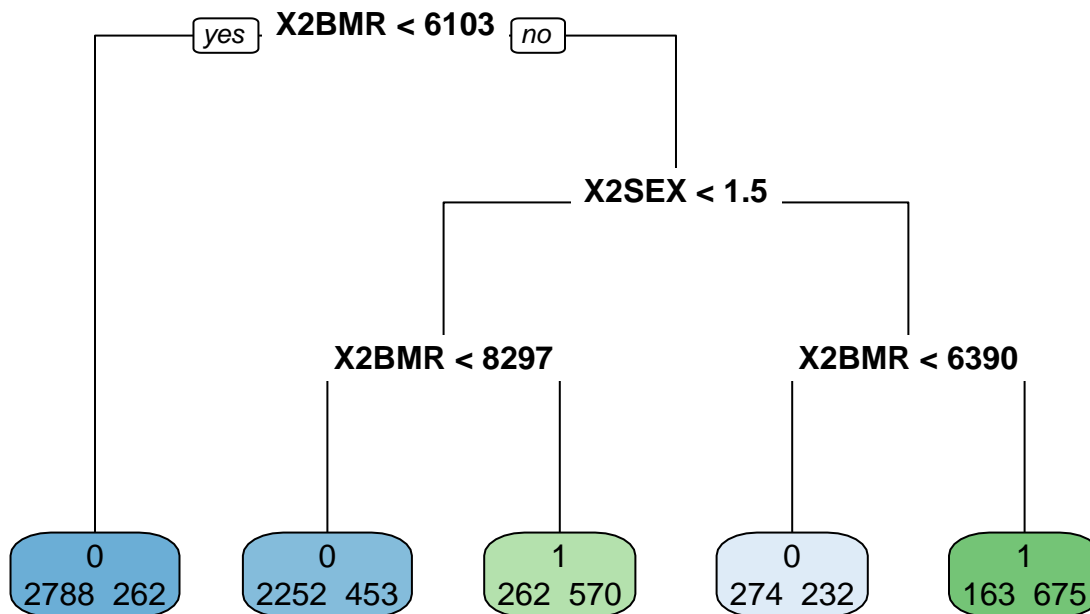
CART Fit for Obesity

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%

Comment:

- 91.4 percent of people(including male and female) whose BMR are less than 6103 are normal (not obese).
- 83.3 percent of male whose BMR are between 6103 and 8297 are normal.
- 68.5 percent of male whose BMR are greater than 8297 are obese
- 54.2 percent of female whose BMR are between 6103 and 6390 are normal
- 80.1 percent of female whose BMR are greater than 6390 are obese

cross-validation for rpart.

CV error

```
res.rpart = cv.rpart(X2,y,V,seed=1)
res.rpart
```

```
## [1] 0.1755138
```

Comment: The CV error for CART is 17.551 percent

Logistic regression

Fit a logistic regression model on most of the data.

```
res.glm = glm(y~.,family=binomial,data=X1)
summary(res.glm)
```

```
##
## Call:
## glm(formula = y ~ ., family = binomial, data = X1)
##
## Deviance Residuals:
##     Min       1Q    Median       3Q      Max
## -2.7799  -0.6546  -0.3851   0.4044   3.2858
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.777e+01  6.188e-01 -28.717  < 2e-16 ***
## BMR          1.879e-03  5.203e-05  36.109  < 2e-16 ***
## EIBMR1      -7.113e-01  7.220e-02  -9.852  < 2e-16 ***
## ADTOTSE     -5.952e-05  2.449e-05  -2.430   0.0151 *
## SEX          3.423e+00  1.174e-01  29.159  < 2e-16 ***
## BDYMSQ04     3.802e-01  8.342e-02   4.557 5.18e-06 ***
## CHOPER1     -1.014e-02  3.948e-03  -2.567   0.0103 *
## FATPER1      1.507e-03  4.357e-03   0.346   0.7295
## PROPER1     -9.982e-04  5.854e-03  -0.171   0.8646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9350.7  on 7930  degrees of freedom
## Residual deviance: 6587.0  on 7922  degrees of freedom
```

```
## AIC: 6605
##
## Number of Fisher Scoring iterations: 5
```

Comment:

The full model has the coefficeints for BMR, EIBMR1, ADTOTSE, SEX, BDYMSQ04, CHOPER1 as statisticlly signifficantly different from zero at the 0.05 level. The fitted model is

$$\text{logit}(p) = -0.1093 + 0.001879 \cdot \text{BMR} - 3.423 \cdot \text{SEX} - 0.7113 \cdot \text{EIBMR1} - 0.00005952 \cdot \text{ADTOTSE} + 0.3802 \cdot \text{BDYMSQ04} - 0.01014 \cdot \text{CH}$$

where p is the probability of obesity.

CV error for glms.

```
## [1] 0.1777834
```

Comment: The CV error for glm is 17.78 percent

summary

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

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

Table 4: Summary of the CV Errors

Methods	Errors
LDA	18.232
CART	17.551
Logistic Regression	17.778