The Effects of age, Gender (male or female), sleeping time, smoking status and employment status on BMI index (Body Mass index measurement).

By: Mostafa Ragheb and Akimawe Kadiri

Instructor's Name: Kevin Foster

Course: B 2000 Statistics and Econometrics

We use the part of BRFSS data from the class website.

The dependent variable Y is BMI index (BMI_measure)

The independent variables X are:

Age, Sleeping time, Smoking status, Gender, Employment Status.

Introduction

Body mass index (BMI) is the cornerstone of the current classification system for obesity and its advantages are widely exploited across disciplines ranging from international surveillance to individual patient assessment. The increased prevalence of obesity and of BMI index in the United States, stresses the pressing need for answers as why this rapid rise has occurred and to know what that the factors that influence greatly the BMI index.

Body Mass Index (BMI) is a measurement of body fat based on height and weight that applies to adult men and women. The BMI formula is simple the mass of a person divided by the square of the height and is defined using the metrics units of kilograms for mass and meters for height. Body Mass Index is an important indicator the health care practitioners to diagnostic if a person overweight or not. A BMI of 25.0 or more is overweight, while the healthy range is 18.5 to 24.9. Many factors contribute and influence the rise and the fall of the Body mass index. Our objective of this paper is to determine what are those factors and ultimately figure out which ones contribute more to the rise of BMI index and obesity in the society. We choose five factors that we think influence the BMI index positively or negatively. We use variables like age (18 to 80), the sleep time, smoking status, gender, and employment status to gauge the body Mass index. We will focus on these factors and try to find which ones have more influence on BMI than the others.

Literature

Previous paper, like the one title “Smoking status and body mass index: A longitudinal study”, the authors investigated the relationship between smoking status and change in body mass index (BMI). They run numerous regressions and statistical comparisons. Their results shows that never- smokers and ex- smokers differ in BMI from current smokers by an average of 1.6kg/m^2. Moreover, their results reveal that smoking increase an BMI of 1.6 kg/m^2. Another article named “Short Sleep Duration is Associated with reduced leptin, elevated Ghrelin, and increased Body Mass Index”; The authors studied the relationship between sleep time and Body Mass index .They use logit regression and some simple regression. Their results shows that participants with short sleep had reduced Leptin and elevated Ghrelin. These differences in Leptin and Ghrelin are likely to increase appetite, explaining the increase BMI observed with short sleep duration. In a nutshell, their study shows that the short sleep increases appetite which leads to an increased BMI. Their research establish a correlation between the number of hours of sleep with high BMI index. An article “A Twin Study of Sleep Duration and Body Mass Index “written by Nathaniel F. Watson M.D study Sleep Duration and Body Mass Index. Their results shows that short sleep was associated with elevated BMI following carefully adjustment for genetics and shared environment. Finally, a research paper “Body Mass Index and Employment Status: Anew look” written by Jonas MinetKinge discuss about the relation between employment status and body mass index. His results shows that factors other than health may be less important in explaining the impact of BMI on employment. He concludes that there is a very weak relationship and correlation between employment status and BMI index.

Methodology

The first regression we run is a simple regression OLS with BMI as y dependent variable and other variables X as independent. We also run a simple regression with BMI square as independent and Age, Gender, and so on as independent variables. we run an OLS with log BMI as dependent variables and with two independent variables Age, Gender. We also use log BMI as dependent variable with sleep time an independent to effectively change from a unit change to a percent change. We run log BMI as dependent with Age as independent variable.

We also run a simple regression as BMI dependent variable with two variables sleep time, Age. We also run log BMI as dependent variable with sleep time, Age as independent variables. We use the logit BMI model to model the probability some dependent variables impact on the BMI. We finally use the RandForest model to classify and to predict of individual X variables.

Results, Analysis and interpretation

Table1

```
lm(formula = BMI ~ Age + Gender + Sleeptime + Smoker + use.to.smoke +
    Never.smoked + Self.employed + Out.of.work.for.a.year.or.more +
    Out.of.work.less.than.a.year + Homemaker + student + Retired +
     Unable.to.work + insufficent.active + Inactive, data = Data_Final)

Residuals:
    Min      1Q  Median      3Q     Max
-16.714  -4.781  -1.150   3.627  43.295

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     29.828132   0.935471  31.886  < 2e-16 ***
Age                             -0.000785   0.015873  -0.049  0.96056
Gender                          -0.233477   0.313025  -0.746  0.45581
Sleeptime                       -0.250623   0.083697  -2.994  0.00278 **
Smoker                          -0.204896   0.609381  -0.336  0.73672
use.to.smoke                     1.733278   0.441451   3.926 8.85e-05 ***
Never.smoked                     0.917576   0.385496   2.380  0.01737 *
Self.employed                   -0.526685   0.548793  -0.960  0.33729
Out.of.work.for.a.year.or.more   2.277256   0.693070   3.286  0.00103 **
Out.of.work.less.than.a.year     1.477426   0.752047   1.965  0.04957 *
Homemaker                       -0.700103   0.466323  -1.501  0.13339
student                         -0.186782   0.823270  -0.227  0.82054
Retired                          1.294531   1.036038   1.250  0.21159
Unable.to.work                   2.106022   0.404651   5.205 2.09e-07 ***
insufficent.active              -0.696698   0.355231  -1.961  0.04995 *
Inactive                        -0.171749   0.326440  -0.526  0.59885
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.906 on 2645 degrees of freedom
Multiple R-squared:  0.02955,    Adjusted R-squared:  0.02405
F-statistic: 5.369 on 15 and 2645 DF,  p-value: 8.508e-11
```

```
cor(Data_Final$BMI , Data_Final$Age)                    0.03389786
```

```
cor(Data_Final$BMI , Data_Final$Sleeptime)              -0.06646048
```

Analysis & interpretation: Table1&Fig1

we run a Linear regression with BMI as a dependent variable y. The table1 shows that the variables “use to smoke” and “Unable to work” are significant. The variable sleep time is less significant in this model.

The Adjusted R square is 0.024, it means that the independent variables explain 2.4% of the target variable. We notice that there is a positive correlation between BMI and Age. Also, it shows that there is a negative correlation between BMI and Sleep time. This result shows that sleep time and Age influence on the BMI index is negligeable.

Fig 1.1 and Fig 1.2 show that are the sleep time has a strong relationship with the BMI. We also observe a tightly clustered around the regression line.

Fig 1.1

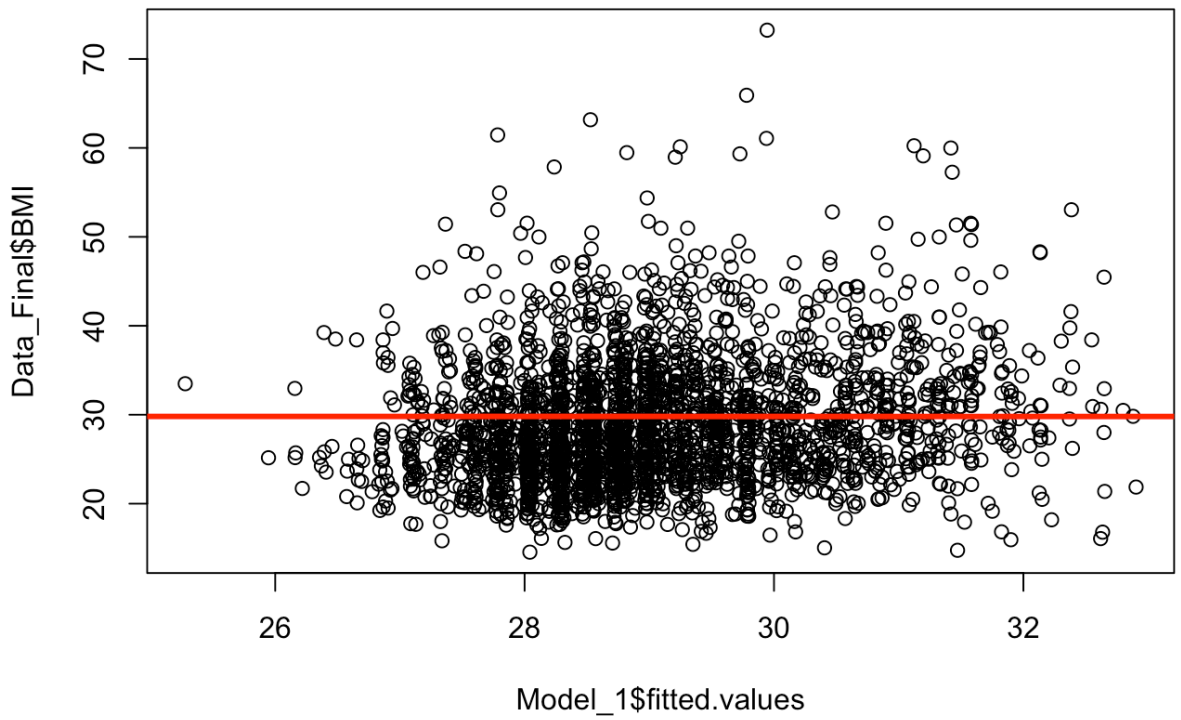

Data_Final$BMI
20
30
40
50
60
70
26
28
30
32
Model_1$fitted.values

Fig 1.2

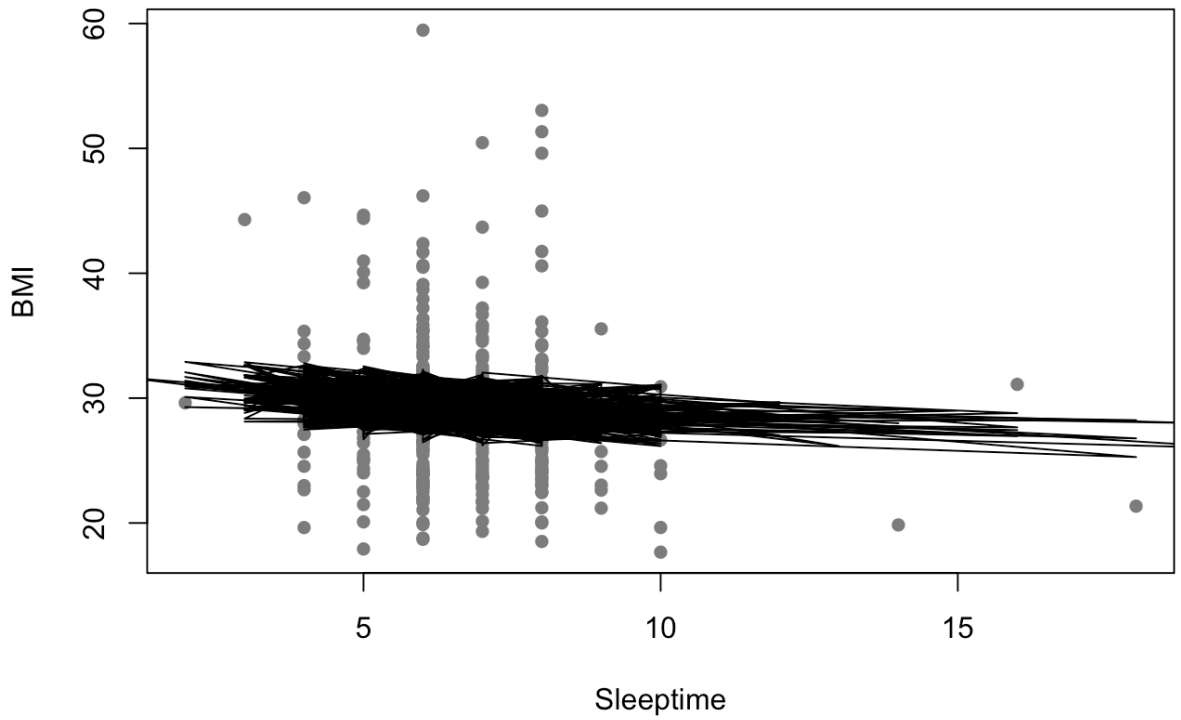


Table2

```
lm(formula = (BMI^2) ~ Age + Gender + Sleeptime + Smoker + use.to.smoke +
       Never.smoked + Self.employed + Out.of.work.for.a.year.or.more +
       Out.of.work.less.than.a.year + Homemaker + student + Retired +
       Unable.to.work + insufficent.active + Inactive, data = Data_Final)

   Residuals:
      Min     1Q Median     3Q    Max
    -860.3 -299.1 -107.0  176.8 4372.8

    Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
   (Intercept)                   946.6447    63.1019  15.002  < 2e-16 ***
   Age                            -0.6561     1.0707  -0.613  0.54008
   Gender                         -4.3930    21.1150  -0.208  0.83521
```

```
Sleeptime                          -15.7450     5.6457  -2.789  0.00533 **
Smoker                             -10.5517    41.1056  -0.257  0.79743
use.to.smoke                       119.3205    29.7779   4.007 6.32e-05 ***
Never.smoked                        64.2884    26.0035   2.472  0.01349 *
Self.employed                      -27.7408    37.0187  -0.749  0.45370
Out.of.work.for.a.year.or.more     147.4064    46.7508   3.153  0.00163 **
Out.of.work.less.than.a.year        96.3613    50.7292   1.900  0.05760 .
Homemaker                          -37.4748    31.4557  -1.191  0.23362
student                             19.3638    55.5335   0.349  0.72735
Retired                             75.3685    69.8857   1.078  0.28093
Unable.to.work                     153.6208    27.2956   5.628 2.01e-08 ***
insufficent.active                 -36.6029    23.9620  -1.528  0.12675
Inactive                            -8.3358    22.0200  -0.379  0.70505

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 465.8 on 2645 degrees of freedom
Multiple R-squared:  0.02769,    Adjusted R-squared:  0.02218
F-statistic: 5.022 on 15 and 2645 DF,  p-value: 7.089e-10
```

Analysis & interpretation: Table 2 & Fig 2.1, Fig2.2

Table2 shows that the variables “use- to- smoke” and “unable -to -work” are significant . the variables Age, Gender, smoke is not significant. we notice that there is a tightly clustered around the regression line between 20 to 40 BMI index in the scatterplot. This means that between that interval the regression shows how well the predicted values fitted in the data.

Fig2.1

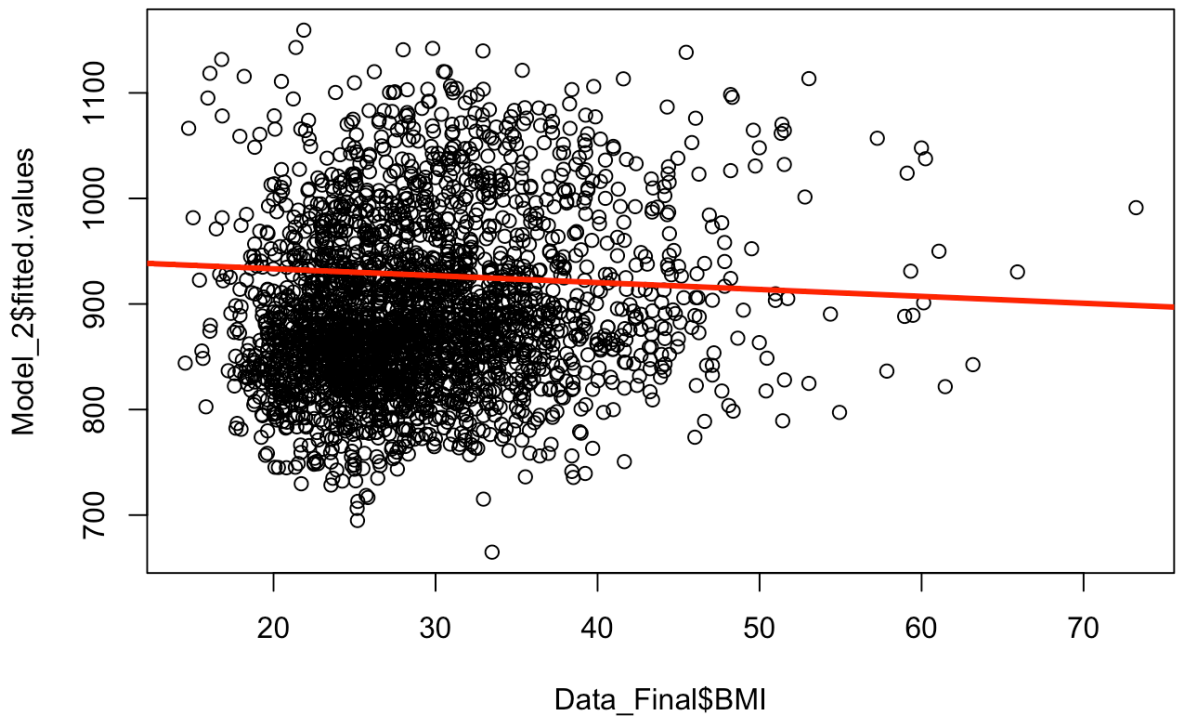


Table3

```
lm(formula = log(BMI) ~ Age + Gender + Sleeptime + Smoker + use.to.smoke +
       Never.smoked + Self.employed + Out.of.work.for.a.year.or.more +
       Out.of.work.less.than.a.year + Homemaker + student + Retired +
       Unable.to.work + insufficent.active + Inactive, data = Data_Final)

   Residuals:
       Min       1Q   Median       3Q      Max
   -0.7202 -0.1565 -0.0161   0.1429   0.9409

   Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
   (Intercept)                       3.3647603  0.0304120 110.639  < 2e-16 ***
```

```
Age                                0.0003269  0.0005160   0.634 0.526433
Gender                            -0.0147370  0.0101764  -1.448 0.147693
Sleeptime                         -0.0083423  0.0027210  -3.066 0.002192 **
Smoker                            -0.0072095  0.0198109  -0.364 0.715948
use.to.smoke                       0.0530358  0.0143515   3.695 0.000224 ***
Never.smoked                       0.0275375  0.0125324   2.197 0.028086 *
Self.employed                     -0.0198163  0.0178412  -1.111 0.266798
Out.of.work.for.a.year.or.more     0.0723349  0.0225316   3.210 0.001342 **
Out.of.work.less.than.a.year       0.0475372  0.0244490   1.944 0.051960 .
Homemaker                         -0.0270605  0.0151601  -1.785 0.074379 .
student                           -0.0192804  0.0267644  -0.720 0.471359
Retired                            0.0448140  0.0336815   1.331 0.183460
Unable.to.work                     0.0589620  0.0131552   4.482 7.71e-06 ***
insufficent.active                -0.0268540  0.0115485  -2.325 0.020130 *
Inactive                          -0.0067004  0.0106125  -0.631 0.527856

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2245 on 2645 degrees of freedom
Multiple R-squared:  0.03025,    Adjusted R-squared:  0.02475
F-statistic:   5.5 on 15 and 2645 DF,  p-value: 3.805e-11
```

Analysis & interpretation: Table 3& Fig 3

We used “log BMI” as dependent variable. Table 1 shows that “use to smoke” and “Unable to work” are significant .The Adjusted R square is 0.024 ,it means that the independent variables explain 2.4% of the target variable.

The Fig3 shows a strong linear positive relationship and there is correlation between the fitted values and BMI index. The slope of the regression line is upward, and it means a positive correlation between BMI and the significant variables. The regression line is not well fitted in the scatterplot.

Fig 3

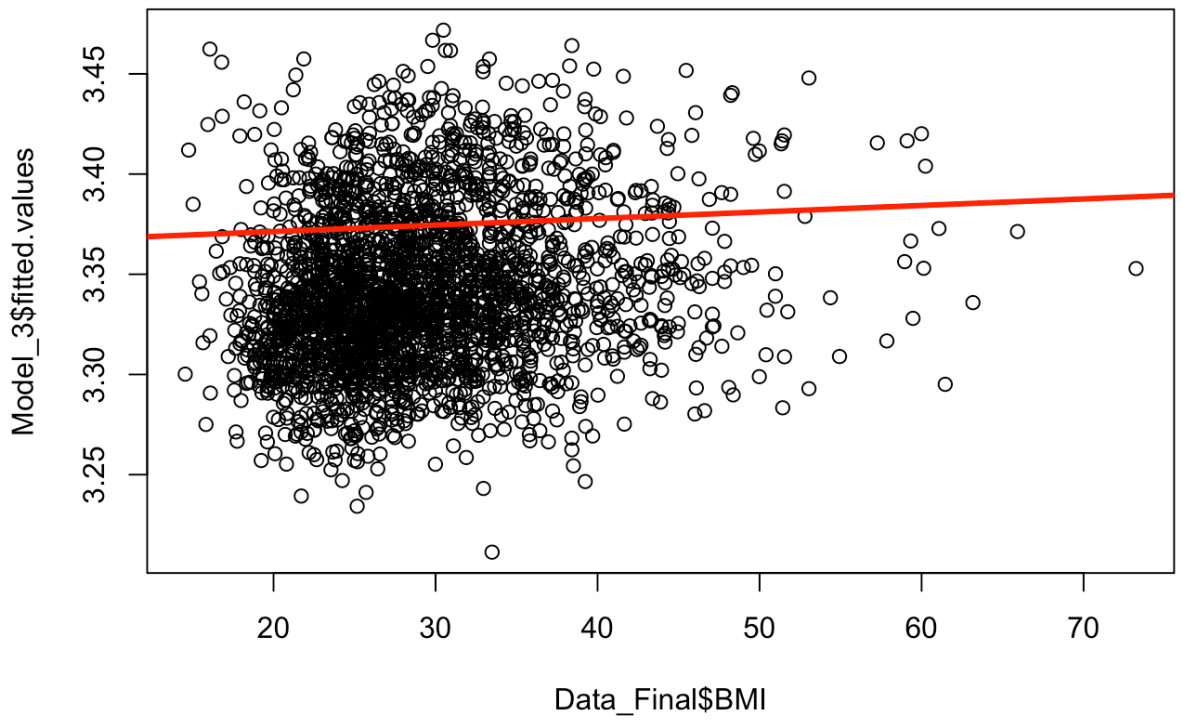


Table 4

```
lm(formula = log(Data_Final$BMI) ~ Data_Final$Sleeptime)

Residuals:
     Min       1Q   Median       3Q      Max
-0.66089 -0.15614 -0.01447  0.14684  0.92701

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.404592   0.018257 186.486  < 2e-16 ***
Data_Final$Sleeptime -0.009466   0.002697  -3.509 0.000457 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2268 on 2659 degrees of freedom
Multiple R-squared:  0.00461,    Adjusted R-squared:  0.004236
```

```
F-statistic: 12.31 on 1 and 2659 DF,  p-value: 0.0004569
```

Analysis & Interpretation: Table 4 &Fig 4

Table4 shows that sleeping time is very significant when we run it as a single independent variable with the BMI as a dependent variable.

Fig 4 shows that it is the best fitted value. This is also showing that people that sleep less than 10 hours a day tend to have a higher BMI index which means that they are obese. Fig 4 shows that there is a strong relationship and a correlation between sleep time and BMI.

Fig 4

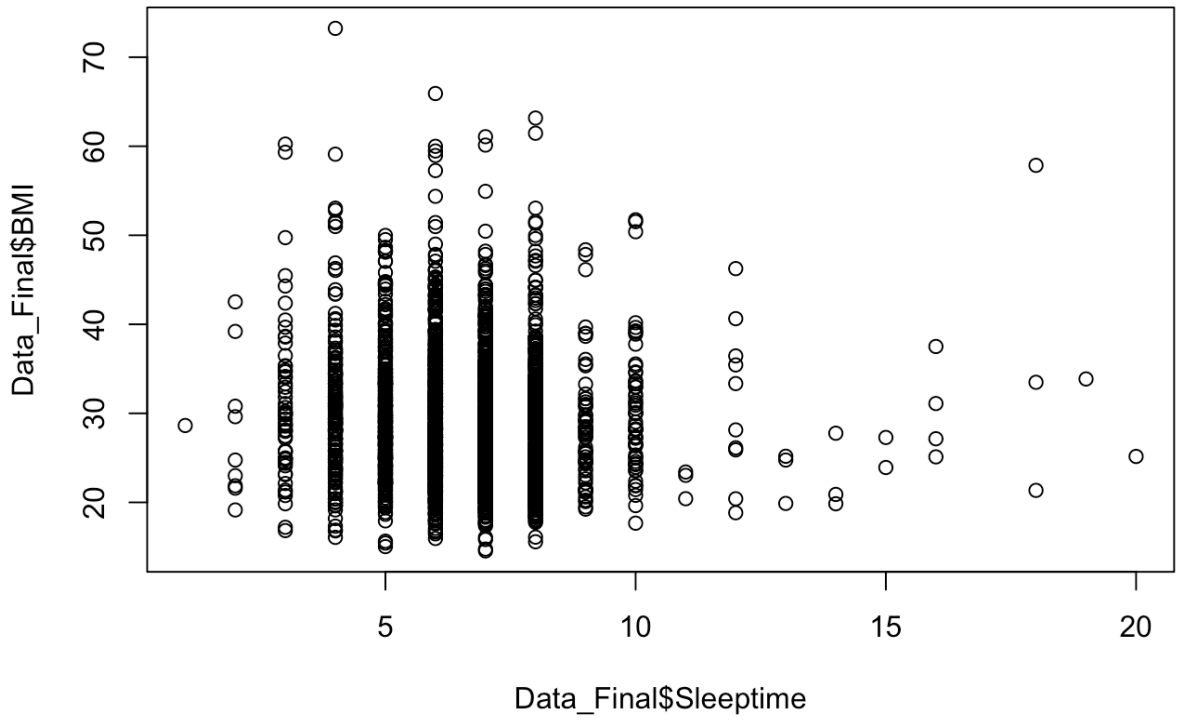


Table 5

```
lm(formula = log(Data_Final$BMI) ~ Data_Final$Age)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-0.65385 -0.15738 -0.01656  0.14550  0.96998

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.2897229  0.0214870 153.103   <2e-16 ***
Data_Final$Age 0.0012153  0.0004851   2.505   0.0123 *

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2271 on 2659 degrees of freedom
Multiple R-squared:  0.002355,   Adjusted R-squared:  0.00198
F-statistic: 6.277 on 1 and 2659 DF,  p-value: 0.01229
```

Analysis & interpretation: Table 5 & Fig 5

Table 5 shows that the variable “Age” is not significant. This indicates that age does not have an effect on the Body Mass Index or obesity.

Fig 5 shows that there is a weak relationship between Age and BMI. Furthermore, we observe that that BMI and Age are uncorrelated. The observations are spread out. This is an indication of uncorrelation.

Fig 5

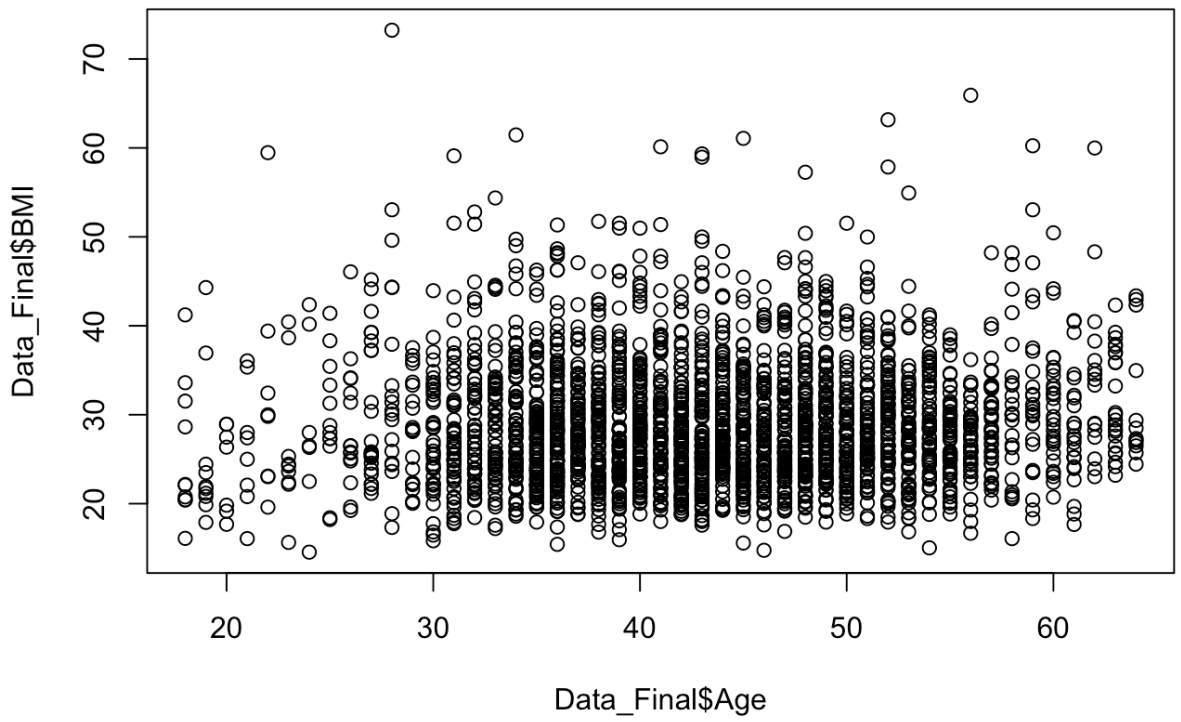


Table 6

```
lm(formula = Data_Final$BMI ~ Data_Final$Sleeptime + Data_Final$Age)

Residuals:
    Min      1Q  Median      3Q     Max
-14.750  -4.852  -1.220   3.622  43.864

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          29.77374    0.85069  34.999  < 2e-16 ***
Data_Final$Sleeptime -0.28667    0.08292  -3.457 0.000555 ***
Data_Final$Age        0.02671    0.01490   1.793 0.073019 .

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.973 on 2658 degrees of freedom
```

```
Multiple R-squared:  0.00562,    Adjusted R-squared:  0.004872
F-statistic: 7.512 on 2 and 2658 DF,  p-value: 0.0005584
```

Analysis and Interpretation: Table 6 & Fig 6

We observe that the variable “Age” is not significant, but the variable “sleep time” is significant. We notice that there is a negative relationship between BMI and sleep time. Also, we observe that there is a negative relationship between BMI and Age. The regression line shows a negative relationship. The regression line is downward sloping which means that there is negative correlation between BMI and sleep time and Age.

Fig 6

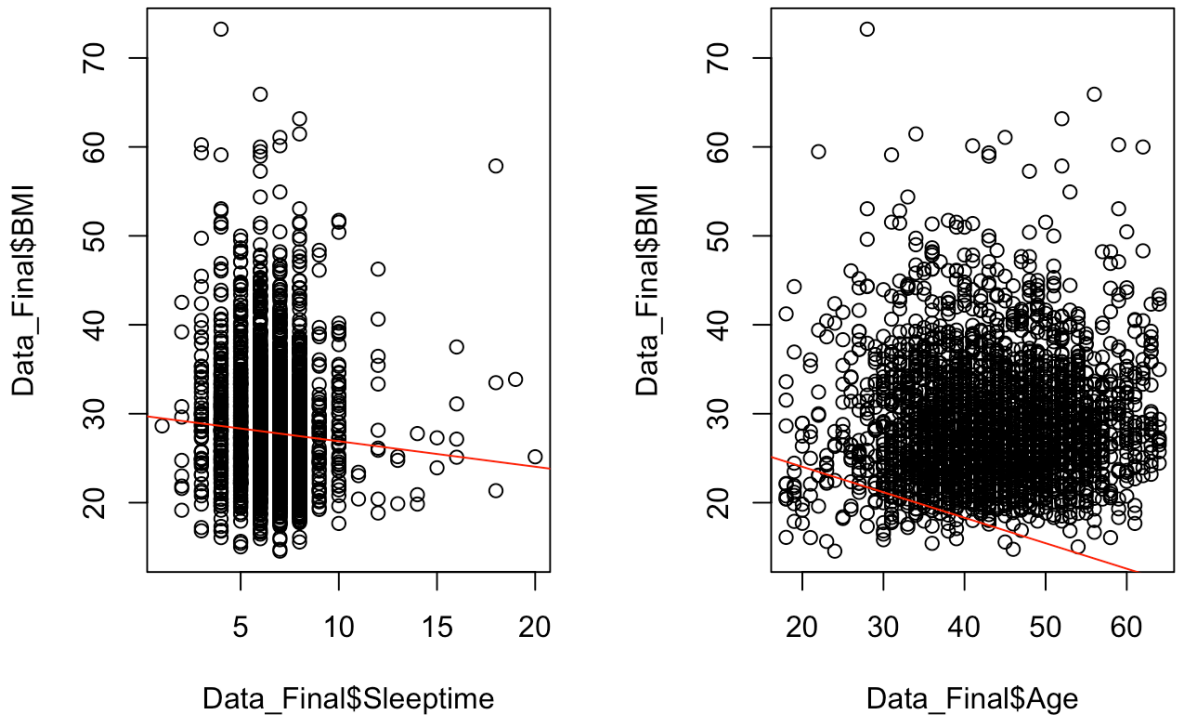


Table 7

```
lm(formula = log(Data_Final$BMI) ~ Data_Final$Sleeptime + Data_Final$Age)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-0.66031 -0.15738 -0.01607  0.14407  0.94577

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.3515571  0.0276460 121.231  < 2e-16 ***
Data_Final$Sleeptime -0.0095473  0.0026948  -3.543 0.000403 ***
Data_Final$Age        0.0012356  0.0004841   2.553 0.010749 *

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2266 on 2658 degrees of freedom
Multiple R-squared:  0.007044,   Adjusted R-squared:  0.006297
F-statistic: 9.428 on 2 and 2658 DF,  p-value: 8.317e-05
```

Analysis and Interpretation: Table 7 & Fig 7

Table 7 shows that “Age” is not significant. On the other hand, “sleeping time” is significant. There is a strong relationship between BMI and sleep time. Fig 7 shows that there is a correlation between sleep time and BMI.

Fig 7

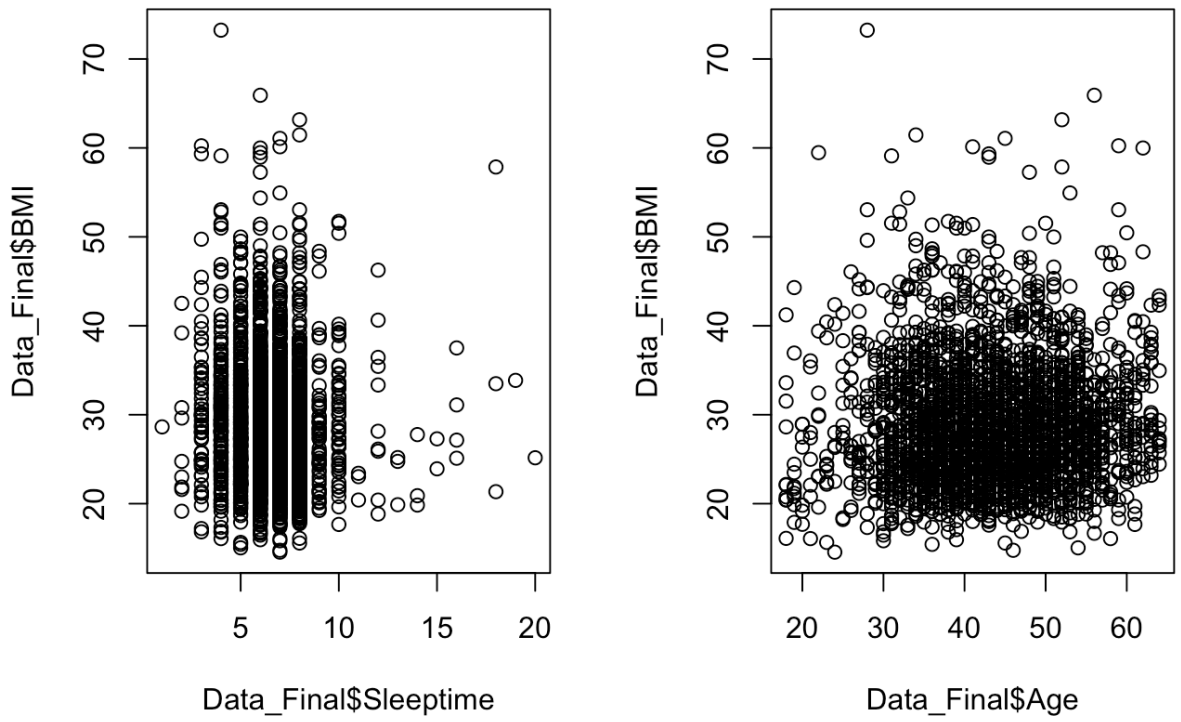


Table 8

```
lm(formula = log(BMI) ~ Sleeptime + use.to.smoke +
Out.of.work.for.a.year.or.more +
       Homemaker + insufficent.active, data = Data_Final)

    Residuals:
       Min      1Q  Median      3Q     Max
   -0.6824 -0.1558 -0.0155  0.1449  0.8923

    Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
   (Intercept)                      3.402257   0.018302 185.900  < 2e-16 ***
   Sleeptime                       -0.009080   0.002693  -3.372 0.000757 ***
   use.to.smoke                     0.035505   0.010605   3.348 0.000825 ***
   Out.of.work.for.a.year.or.more   0.056642   0.022126   2.560 0.010524 *
   Homemaker                       -0.042227   0.014506  -2.911 0.003633 **
```

```
  insufficent.active             -0.029425   0.010914  -2.696 0.007062 **

  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  Residual standard error: 0.2255 on 2655 degrees of freedom
  Multiple R-squared:  0.01747,    Adjusted R-squared:  0.01562
  F-statistic: 9.441 on 5 and 2655 DF,  p-value: 6.189e-09
  Mode   FALSE    TRUE
logical    2386     275
```

Analysis and Interpretation: Table 8 & Fig 8

Table 8 shows that “Sleep time” and “use to smoke” are significant. This means that a short sleep time has a negative effect on the Body mass index. In other word, the less you sleep, the greater your BMI index will be. This will lead to obesity.

We also observe that the variable “use to smoke” has a positive effect on the BMI index. Table8 shows that people that use to smoke have low Body mass index.

Table 8 indicate that the Mode FALSE being 2386 means there is a strong correlation between sleep time, use to smoke with body mass index.

We observed in Fig 8 that BMI is concentrated between the 3.30 and 3.40 which means that the values are well and best fitted.

Fig 8

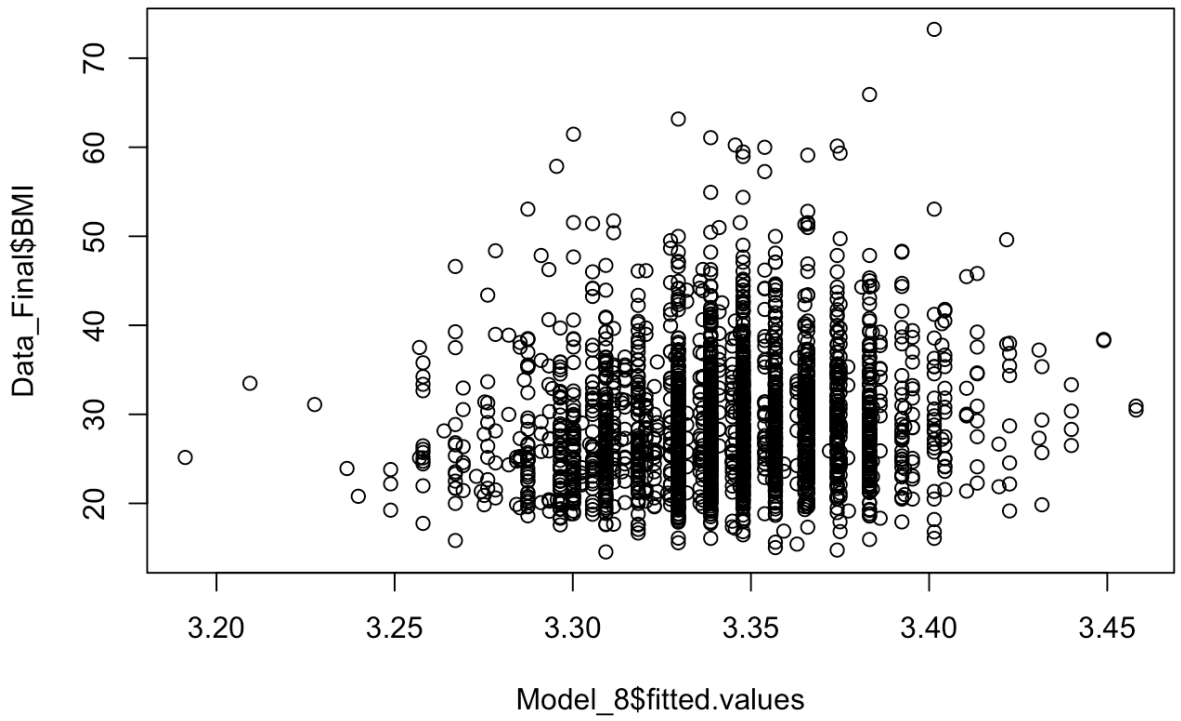


Table 9

```
Age.V1         Gender       Sleeptime.V1      Smoker    use.to.smoke
    Min.   :-2.7434434   1:1763   Min.   :-3.268136   1: 156   1: 522
    1st Qu.:-0.7279074   0: 623   1st Qu.:-0.274341   0:2230   0:1864
    Median :-0.0914224            Median : 0.324419
    Mean   :-0.0599005            Mean   : 0.073974
    3rd Qu.: 0.6511435            3rd Qu.: 0.324419
    Max.   : 2.1362753            Max.   : 8.108287
    Never.smoked Self.employed Out.of.work.for.a.year.or.more
    1:1285       1: 158        1: 100
    0:1101       0:2228        0:2286



    Out.of.work.less.than.a.year Homemaker student  Retired  Unable.to.work
```

```
 1:  78                        1: 246    1:  72   1:  43   1: 390
 0:2308                        0:2140    0:2314   0:2343   0:1996



 insufficent.active Inactive       BMI.V1
 1: 484             1: 614   Min.   :-1.922332
 0:1902             0:1772   1st Qu.:-0.685329
                           Median :-0.212601
                              Mean   :-0.062216
                           3rd Qu.: 0.398891
                              Max.   : 5.632347
```

Analysis and interpretation: Table 9

Table 9 shows that the min. of Gender and smoker are negative. But the min. of sleep time is positive. We also observe that the variable use-to-smoke is relevant and has a min. which is positive.

Table 10

```
lm(formula = sobj$formula, data = sobj$data)

   Residuals:
        Min       1Q   Median       3Q      Max
   -1.13487 -0.29222  0.05445  0.27737  0.67001

   Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
   (Intercept)                       0.730781   0.171119   4.271 2.74e-05 ***
    Age                               0.035218   0.022333   1.577   0.1160
   Gender1                          -0.040989   0.025436  -1.611   0.1083
   Sleeptime                        -0.050475   0.021326  -2.367   0.0187 *
   Smoker1                          -0.068161   0.045159  -1.509   0.1324
   use.to.smoke1                     0.003055   0.035740   0.085   0.9319
   Never.smoked1                     0.030911   0.030555   1.012   0.3126
   Self.employed1                    0.009980   0.040232   0.248   0.8043
   Out.of.work.for.a.year.or.more1 -0.047912   0.060091  -0.797   0.4260
```

```
Out.of.work.less.than.a.year1    0.057943   0.050217   1.154   0.2496
Homemaker1                       0.064076   0.037814   1.695   0.0914 .
student1                         0.001996   0.072424   0.028   0.9780
Retired1                         0.019055   0.079603   0.239   0.8110
Unable.to.work1                  0.022513   0.031336   0.718   0.4731
insufficent.active1             -0.020315   0.028093  -0.723   0.4703
Inactive1                       -0.031857   0.025792  -1.235   0.2179
BMI                              0.302981   0.021269  14.245  < 2e-16 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3413 on 258 degrees of freedom
Multiple R-squared:  0.4847, Adjusted R-squared:  0.4528
F-statistic: 15.17 on 16 and 258 DF,  p-value: < 2.2e-16
      True
pred       0    1
  FALSE  580  126
  TRUE   161 1519
```

Analysis and Interpretation: Table 10

Table 10 is a regression between “Overweight and all other variables” showing that the explanation of obesity is only by BMI. The variable sleep time is very weakly significant in this regression.

Table 10 is showing that the model predicted People Not overweight are 580 with error of 126, giving the percentage error equal: 17.8% While people overweight are 1519 with an error of 161, giving the percentage error equal 9.5%

```
126/(580+126)                    0.1784703

161/(1519+161)                   0.09583333
```

Table 11

```
glm(formula = sobj$formula, family = binomial, data = sobj$data)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-1.711e-04  -2.100e-08   2.100e-08   2.100e-08   2.013e-04

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                        26.658 184039.932   0.000    1.000
Age                                 2.603   4723.874   0.001    1.000
Gender1                           -17.158   3062.813  -0.006    0.996
Sleeptime                         -15.705   3312.528  -0.005    0.996
Smoker1                             7.971   5091.985   0.002    0.999
use.to.smoke1                      -1.702  18023.044   0.000    1.000
Never.smoked1                      29.380   5532.607   0.005    0.996
Self.employed1                     -5.882  25173.348   0.000    1.000
Out.of.work.for.a.year.or.more1  -129.707  75234.384  -0.002    0.999
Out.of.work.less.than.a.year1     -10.437  84924.724   0.000    1.000
Homemaker1                          5.224   5398.160   0.001    0.999
student1                          -24.829  59187.792   0.000    1.000
Retired1                          -73.561 129223.658  -0.001    1.000
Unable.to.work1                     4.566   5233.999   0.001    0.999
insufficent.active1                -2.647   5079.803  -0.001    1.000
Inactive1                         -17.560   4048.405  -0.004    0.997
BMI                               458.611  43368.236   0.011    0.992

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3.3848e+02  on 274  degrees of freedom
Residual deviance: 2.4994e-07  on 258  degrees of freedom
AIC: 34

Number of Fisher
       true
```

```
pred       0    1
  FALSE  705   86
  TRUE    36 1559
```

Analysis and Interpretation: Table 11

Since a high BMI meaning Obesity, GLM model is used to measure the effect of the variables being used on obesity, the variables Age, Gender, use to smoke and sleep time have the negative coefficients. It can be interpret as having less effect on Obesity. On the other hand, the variable smoker has a positive coefficient. This means that smokers increase Obesity. This shows that the BMI and smoker are correlated. The Akaike Information Criterion (AIC) in Table 11 is 34 and low. This means that it is a good model, and we have a better fit.

Table 1 1 is showing that the model predicted people not overweight are 705 with error of 86, giving the percentage error 10.8 % while people overweight are 1559 with an error of 36, giving the percentage error equal 2.25%.

```
86/(705+86)                             0.1087231
36/(36+1559)                            0.02257053
```

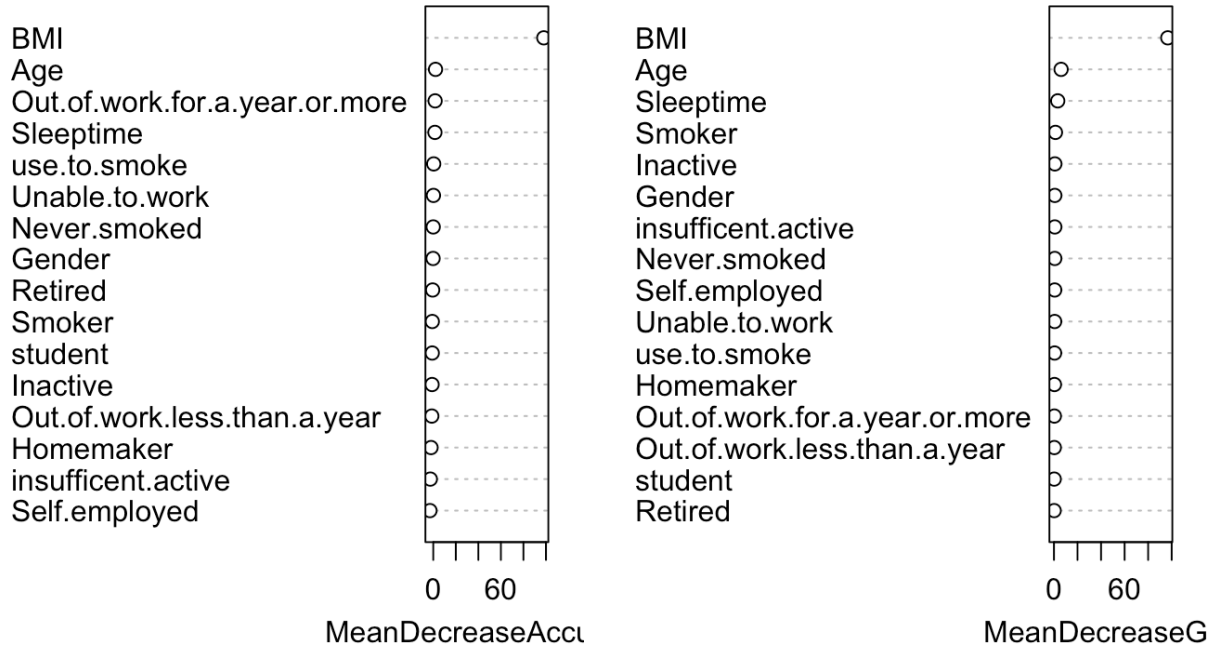
Table 12
Randomforest

```
randomForest(formula = as.factor(Overweight) ~ ., data = sobj$data,
importance = TRUE, proximity = TRUE)
                  Type of random forest: classification
                        Number of trees: 500
   No. of variables tried at each split: 4

           OOB estimate of  error rate: 1.45%
   Confusion matrix:
      0   1 class.error
   0 83   1  0.01190476
   1  3 188  0.01570681
```

```
     0     1 MeanDecreaseAccuracy
```

```
                                  0.98  1.79                2.01
Gender                           -1.63  1.11               -0.16
Sleeptime                         2.39 -0.20                1.47
Smoker                           -1.71  0.28               -0.86
use.to.smoke                      0.01  0.41                0.40
Never.smoked                     -0.56  0.05               -0.02
Self.employed                    -1.94 -2.39               -2.92
Out.of.work.for.a.year.or.more    3.39 -0.51                1.65
Out.of.work.less.than.a.year     -0.46 -1.40               -1.30
Homemaker                        -0.03 -3.08               -1.97
student                          -0.51 -0.52               -0.90
Retired                           0.30 -1.40               -0.54
Unable.to.work                    0.70 -0.74                0.17
insufficent.active               -2.49 -1.41               -2.60
Inactive                         -2.68  0.69               -1.05
BMI                              96.89 87.55               98.06
                                 MeanDecreaseGini
Age                                          6.06
Gender                                       0.72
Sleeptime                                    3.20
Smoker                                       1.28
use.to.smoke                                 0.54
Never.smoked                                 0.68
Self.employed                                0.62
Out.of.work.for.a.year.or.more               0.33
Out.of.work.less.than.a.year                 0.29
Homemaker                                    0.41
student                                      0.20
Retired                                      0.16
Unable.to.work                               0.59
insufficent.active                           0.70
Inactive                                     0.76
BMI                                         96.69
```




```
      true
pred     0    1
   0   737    8
   1     4 1637
```

Analysis & interpretation: Forest model

The mean decrease Gini shows that the Gini index of node impurity decrease with the variables. We observe that BMI, Age, Gender, sleep time and smoker are at the top of the mean decrease Gini classification. This means that Age, Gender, sleep time and smoker have more effect on BMI than other variables. BMI has 96.69 which is the higher number follow by Age, sleep time, smoker and Gender.

According to forest model the error percent is predicted people who are not overweight are 0.01190476 and predicted people who are overweight are 0.01570681. This means that prediction is accurate with the OBB estimate error rate of 1.45% in the model.

Conclusion

Our findings show that even though Aging contribute a little bit to the increase of the body Mass index, but it is negligible compare to the other factor. The relationship between Age and BMI is relative. Our research shows in most of the model that sleep time is very correlated to the BMI index and consequently to the obesity. We observed that the shorter the duration of an adult is, the greater is his BMI index. All our models classify the variable sleep time as one of the factors that impact the BMI index. People that sleep less than 5hours a day tend to have a higher body mass index. But people that sleep an average of 7 to 9 hours tend to have for the most part a low BMI index and consequently less prone to obesity. We notice that the variable use-to-smoke is also one of the factors that has a huge influence on the Body mass index. Most of the regressions we run, show that the variable use-to-smoke is significant. The variable use-to-smoke is correlated. Our research reveal that the variable use-to-smoke has a negative impact on BMI. This means that a person that use to smoke has a higher BMI index than a person who still smoking. Most people tend to gain a lot of weight as soon as the stop smoking, and this leads to an increase of people Body Mass index. The increased of BMI index inevitably leads to obesity which is a serious health problem. Another one of the factors that has an impact on BMI index is the variable Unable.to. work. people that are unable to work have a limited mobility and activity. People with different disability, inactive people that are unable to work, most of time tend to have a higher BMI index. The variable unable to work is positively and strongly correlated to BMI.

We observed Gender is insignificant in all our regressions model. This means that being male or female does not increase or decrease your BMI index. The factor Gender is weakly correlated to the body mass index or obesity. Unlike previous papers that find Age and gender as factors that have a great impact on Body Mass index, our experimentation and our observations find variables Sleep- time and use-to-smoke with the greater effect on BMI.

References

Marcus R. Munafo , Kate Tilling , Yoaz Ben-Shlomo Nicotine & Tobacco Research , Volume ll , Issue 6, June 2009,Pages 765 -771, https://doi.org/10.1093/ntr/ntp062

Shahrad Tahen , ling Lin, Diane Austin , Terry Young, Emmanuel Mignot, Short Sleep Duration is Associated with Reduced Leptin , Elevated Ghrelin,and increased Body Mass Index . Publish : December 7, 2004. https://doi.org/10.1371/Journal.pmed .0010062

Nathaniel Watson ,MD , Debra Buchwald MD, Michael V. Vitiello, Phd, Carolyn Noonan, MS ,Jack Goldberg, phd, Published online: February 15,2010. https://doi.org/10.5667/jcsm.27704.

James W. Youdas ,PTMS, John H. Hollman , PT Phd & David A. Krause , DSCPT MBA OCS Pages 229-237/Accepted 11 Nov 2005, Published Online : 10 jul 2009

Beyond body mass index by A.M. Prentice, S.A. Jebb 21 December 2001, https://doi.org/10.1046/J.1467-789X.2001.00031X.