650 Final Project

Descriptive Analysis

Yubo SHAO

2021-11-24

```r
library(NHANES)
raw.data <- force(NHANES) #import dateset

##  only keep "SleepHrsNight" "SleepTrouble", "Age", "Gender",
## "Poverty", "PhysActive", "AlcoholYear"
data.selected <- raw.data %>% select(c("SleepHrsNight", "SleepTrouble", "Age", "Gender",
                                       "Poverty", "PhysActive", "AlcoholYear", "Depressed"))
dim(data.selected)
## [1] 10000     8
```

1.Preliminary data clean

(1) Define study population & check missing values

```r
## Study population: (1) age 18+ (2) No SleepTrouble
data.population <- data.selected[which(data.selected$Age >= 18 & data.selected$SleepTrouble == "No"),]
dim(data.population)
## [1] 5537    8

## missing value
missing.table <- data.frame(
  col1 = c("SleepHrsNight", "Age", "Gender", "Poverty", "Depressed", "PhysActive", "AlcoholYear"),
  col2 = c(sum(is.na(data.population$SleepHrsNight)),sum(is.na(data.population$Age)),
           sum(is.na(data.population$Gender)),sum(is.na(data.population$Poverty)),
           sum(is.na(data.population$Depressed)), sum(is.na(data.population$PhysActive)),
           sum(is.na(data.population$AlcoholYear))),
  col3 = c(sum(!is.na(data.population$SleepHrsNight)),sum(!is.na(data.population$Age)),
           sum(!is.na(data.population$Gender)),sum(!is.na(data.population$Poverty)),
           sum(!is.na(data.population$Depressed)), sum(!is.na(data.population$PhysActive)),
           sum(!is.na(data.population$AlcoholYear))))
missing.table$col4 <- round(missing.table$col3/(missing.table$col2 + missing.table$col3) * 100,
                            digits = 2)
colnames(missing.table) <- c("Variables", "#. missing", "#. completed", "% completed")
print(kable(missing.table, align = c('c', 'c', 'c', 'c')))
##
##
## |   Variables   | #. missing | #. completed | % completed |
## |:-------------:|:----------:|:------------:|:-----------:|
## | SleepHrsNight |     3      |     5534     |    99.95    |
## |      Age      |     0      |     5537     |   100.00    |
```

```
## |    Gender     |     0      |     5537     |   100.00    |
## |    Poverty    |    438     |     5099     |    92.09    |
## |   Depressed   |    603     |     4934     |    89.11    |
## |  PhysActive   |     0      |     5537     |   100.00    |
## |  AlcoholYear  |    1220    |     4317     |    77.97    |
```

(2) Delete missing value

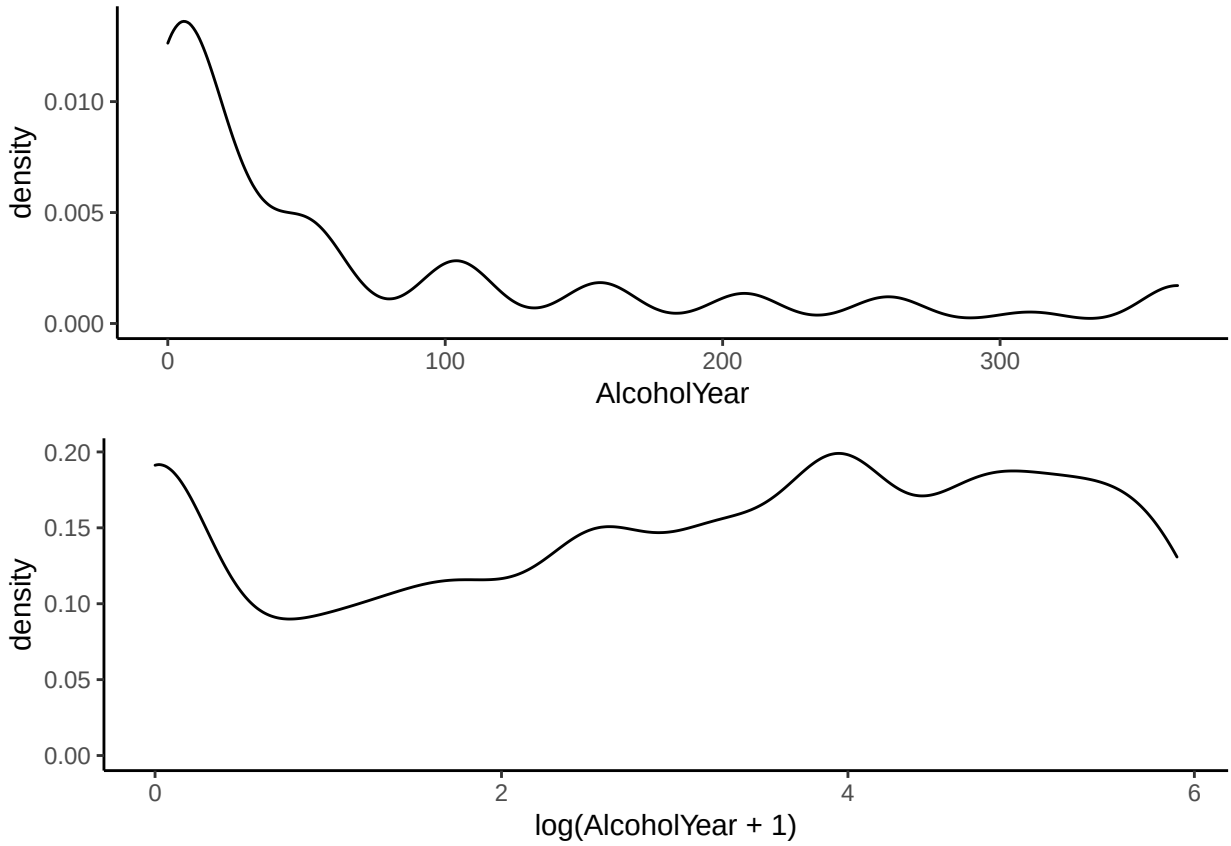
```
data.complete <-  na.omit(data.population)
# dim(data.complete)
# head(data.complete)
```

(3) Create new age groups (by quartile) & combine "Depressed" & centered continuous variable

```
## Create new age groups
quantile.age <- quantile(data.complete$Age)
print(quantile.age)
##   0%  25%  50%  75% 100%
##   18   31   44   58   80
data.complete$Age_cate1 <- 1 * (data.complete$Age < quantile.age[2])  #group1: [18,31)
data.complete$Age_cate2 <- 1 * (data.complete$Age >= quantile.age[2]
                                & data.complete$Age < quantile.age[3]) #group1: [31,44)
data.complete$Age_cate3 <- 1 * (data.complete$Age >= quantile.age[3]
                                & data.complete$Age < quantile.age[4]) #group1: [44,58)
data.complete$Age_cate4 <- 1 * (data.complete$Age >= quantile.age[4])  #group4: [58,80)

## combine "Depressed"
## None = "No"; Several & Most ="Yes"
data.complete$depressed_binary[data.complete$Depressed == "None"] <- "No"
## Warning: Unknown or uninitialised column: `depressed_binary`.
data.complete$depressed_binary[data.complete$Depressed == "Several" |
                               data.complete$Depressed == "Most"] <- "Yes"

## centered Poverty (- median)
## The main reason for data centralization is to make it easier to interpret the results.
median.poverty <- median(data.complete$Poverty)
data.complete$center.poverty <- data.complete$Poverty - median.poverty
```

(4) Do log-transformation on AlcoholYears

```
data.clean <- data.complete %>% select(-c("Age", "SleepTrouble", "Depressed", "Poverty"))
AlcoholYear.plot <- ggplot() + geom_density(aes(log(AlcoholYear+1)), data.clean) +
    theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black"))

AlcoholYear.plot2 <- ggplot() + geom_density(aes(AlcoholYear), data.clean) +
    theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black"))

grid.arrange(AlcoholYear.plot2, AlcoholYear.plot, nrow = 2)
```




```
head(data.clean, n = 5)
## # A tibble: 5 x 10
##   SleepHrsNight Gender PhysActive AlcoholYear Age_cate1 Age_cate2 Age_cate3
##           <int> <fct>  <fct>            <int>     <dbl>     <dbl>     <dbl>
## 1             8 female Yes                 52         0         0         1
## 2             8 female Yes                 52         0         0         1
## 3             8 female Yes                 52         0         0         1
## 4             7 male   Yes                100         0         0         0
## 5             5 male   Yes                104         0         0         0
## # ... with 3 more variables: Age_cate4 <dbl>, depressed_binary <chr>,
## #   center.poverty <dbl>
```

2. Fit the model

(1) SLR: PhysActive v.s. SleepHrsNight

$\beta = 0.094$, and p=0.018

```
lm.model1 <- lm(SleepHrsNight ~ PhysActive)
summary(lm.model1)
##
## Call:
## lm(formula = SleepHrsNight ~ PhysActive)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0959 -1.0023 -0.0959  0.9041  4.9977
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.00227    0.02963 236.304   <2e-16 ***
## PhysActiveYes  0.09364    0.03956   2.367    0.018 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.244 on 4012 degrees of freedom
## Multiple R-squared:  0.001395,   Adjusted R-squared:  0.001146
## F-statistic: 5.603 on 1 and 4012 DF,  p-value: 0.01798
```

(2) MLR with interaction

```
lm.model3 <- lm(SleepHrsNight ~ Gender + center.poverty + PhysActive + log(AlcoholYear + 1) +
                  depressed_binary + Age_cate2 + Age_cate3 + Age_cate4 +
                  PhysActive*Age_cate2 + PhysActive*Age_cate3 + PhysActive*Age_cate4)
summary(lm.model3)
##
## Call:
## lm(formula = SleepHrsNight ~ Gender + center.poverty + PhysActive +
##     log(AlcoholYear + 1) + depressed_binary + Age_cate2 + Age_cate3 +
##     Age_cate4 + PhysActive * Age_cate2 + PhysActive * Age_cate3 +
##     PhysActive * Age_cate4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4028 -0.8192  0.0032  0.8557  5.0019
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                7.323732   0.078944  92.771  < 2e-16 ***
## Gendermale                -0.253049   0.039384  -6.425 1.47e-10 ***
## center.poverty             0.001321   0.012879   0.103 0.918335
## PhysActiveYes             -0.202547   0.084056  -2.410 0.016012 *
## log(AlcoholYear + 1)       0.037955   0.010519   3.608 0.000312 ***
## depressed_binaryYes       -0.279009   0.053528  -5.212 1.96e-07 ***
## Age_cate2                 -0.290844   0.091728  -3.171 0.001532 **
## Age_cate3                 -0.577462   0.089385  -6.460 1.17e-10 ***
## Age_cate4                 -0.010000   0.086439  -0.116 0.907901
## PhysActiveYes:Age_cate2    0.145711   0.115086   1.266 0.205548
## PhysActiveYes:Age_cate3    0.599800   0.113063   5.305 1.19e-07 ***
## PhysActiveYes:Age_cate4    0.293676   0.113787   2.581 0.009889 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.218 on 4002 degrees of freedom
## Multiple R-squared:  0.04449,    Adjusted R-squared:  0.04186
## F-statistic: 16.94 on 11 and 4002 DF,  p-value: < 2.2e-16
m3.residuals <- lm.model3$residuals
```

You can get some ideas about how to create a summary table of MLR from this article (Table3).

(3) calculate point estimate & standard error

```
summary.model3 <- summary(lm.model3)
beta <- summary.model3$coefficients[c(1, 4, 7, 8, 9, 10, 11, 12), 1]
std <- summary.model3$coefficients[c(1, 4, 7, 8, 9, 10, 11, 12), 2]

## Age group1
Phy.no.meanY.in.age1 <-
  c(c(1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0) %*% coef(summary.model3))[1]
Phy.no.std.in.age1 <-
  c(c(1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0) %*% vcov(lm.model3) %*%
      c(1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0))
confint.age1.1 <-
  Phy.no.meanY.in.age1 + c(-1, 1) * qt(p = 0.975, df = summary.model3$df[2]) *
  sqrt(Phy.no.std.in.age1)

Phy.yes.meanY.in.age1 <-
  c(c(1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0) %*% coef(summary.model3))[1]
Phy.yes.std.in.age1 <-
  c(c(1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0) %*% vcov(lm.model3) %*%
      c(1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0))
confint.age1.2 <-
  Phy.yes.meanY.in.age1 + c(-1, 1) * qt(p = 0.975, df = summary.model3$df[2]) *
  sqrt(Phy.yes.std.in.age1)

## Age group2
Phy.no.meanY.in.age2 <-
  c(c(1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0) %*% coef(summary.model3))[1]
Phy.no.std.in.age2 <-
  c(c(1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0) %*% vcov(lm.model3) %*%
      c(1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0))
confint.age2.1 <-
  Phy.no.meanY.in.age2 + c(-1, 1) * qt(p = 0.975, df = summary.model3$df[2]) *
  sqrt(Phy.no.std.in.age2)

Phy.yes.meanY.in.age2 <-
  c(c(1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0) %*% coef(summary.model3))[1]
Phy.yes.std.in.age2 <-
  c(c(1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0) %*% vcov(lm.model3) %*%
      c(1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0))
confint.age2.2 <-
  Phy.yes.meanY.in.age2 + c(-1, 1) * qt(p = 0.975, df = summary.model3$df[2]) *
  sqrt(Phy.yes.std.in.age2)

## Age group3
Phy.no.meanY.in.age3 <-
  c(c(1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0) %*% coef(summary.model3))[1]
Phy.no.std.in.age3 <-
  c(c(1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0) %*% vcov(lm.model3) %*%
      c(1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0))
confint.age3.1 <-
  Phy.no.meanY.in.age3 + c(-1, 1) * qt(p = 0.975, df = summary.model3$df[2]) *
  sqrt(Phy.no.std.in.age3)
```

```r
Phy.yes.meanY.in.age3 <-
  c(c(1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0) %*% coef(summary.model3))[1]
Phy.yes.std.in.age3 <-
  c(c(1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0) %*% vcov(lm.model3) %*%
      c(1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0))
confint.age3.2 <-
  Phy.yes.meanY.in.age3 + c(-1, 1) * qt(p = 0.975, df = summary.model3$df[2]) *
  sqrt(Phy.yes.std.in.age3)

## Age group4
Phy.no.meanY.in.age4 <-
  c(c(1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0) %*% coef(summary.model3))[1]
Phy.no.std.in.age4 <-
  c(c(1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0) %*% vcov(lm.model3) %*%
      c(1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0))
confint.age4.1 <-
  Phy.no.meanY.in.age2 + c(-1, 1) * qt(p = 0.975, df = summary.model3$df[2]) *
  sqrt(Phy.no.std.in.age2)

Phy.yes.meanY.in.age4 <-
  c(c(1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1) %*% coef(summary.model3))[1]
Phy.yes.std.in.age4 <-
  c(c(1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1) %*% vcov(lm.model3) %*%
      c(1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1))
confint.age4.2 <-
  Phy.yes.meanY.in.age4 + c(-1, 1) * qt(p = 0.975, df = summary.model3$df[2]) *
  sqrt(Phy.yes.std.in.age4)
```

(4) Effect of PhyActive in different age groups

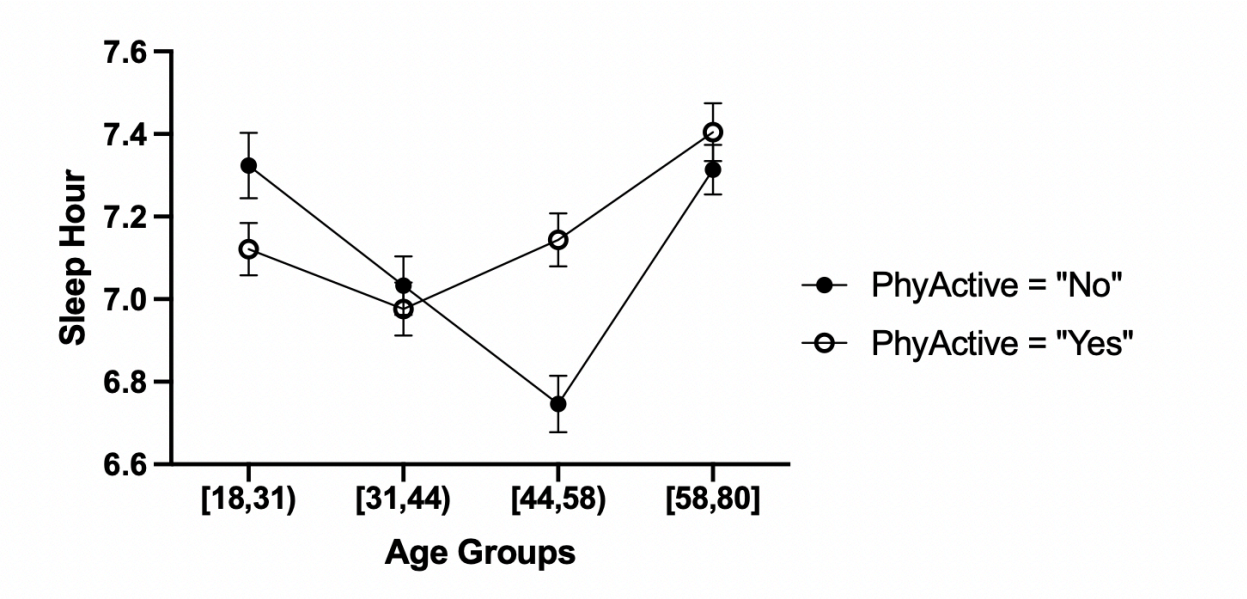


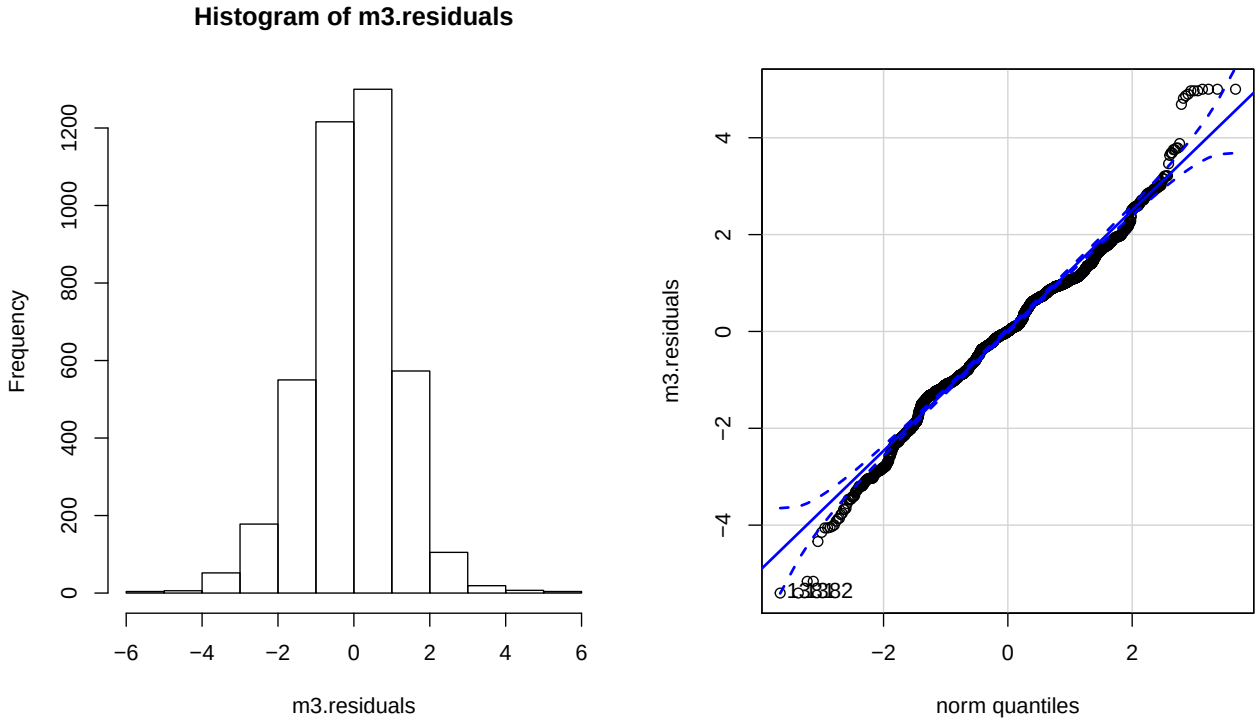
Figure 1: Effect.Figure

I created this figure by *Prism* 9,but we should use R to get this plot

If Runshu and Haoyue are not familiar with ggplot2, you can find some tutorials from this *link*

(5) Check residuals

```
par(mfrow = c(1, 2))
hist(m3.residuals)
car::qqPlot(m3.residuals)
```




```
## [1] 1381 1382
```

(6) Sequential F-test

Check whether age modified the association between PhyActive and Sleep time.

```
anova.table <- anova(lm.model3)
sequential.F.statistics <- ((anova.table$`Sum Sq`[9] + anova.table$`Sum Sq`[10] +
                               anova.table$`Sum Sq`[11])/3)/anova.table$`Mean Sq`[12]
print(sequential.F.statistics)
## [1] 10.59669
qf(0.95, 3, anova.table$Df[12])
## [1] 2.607128
pf(sequential.F.statistics, 3, anova.table$Df[12], lower.tail = FALSE) # p <0.001
## [1] 6.146971e-07
```

3. Brief intepretation:

- We estimated that, among people aged 18-31, comparing people who don't have physical activities, the sleep time of those who have physical activities is 0.203 hours significantly shorter (p=0.016), adjusted for others covariates.

- Age will change the effect of physical activities on sleep hour (Sequential F test, $p < 0.001$).
- Report the point estimate of physical activities effects and p-value in each age group (among female (Gendermale = 0), who has median poverty (center.poverty = 0), never drink (log(AlcoholYear + 1) = 0), no depression (depressed_binaryYes = 0), and no sleep disorder). (need to calculate $se(\sum \beta' s)$)
- ...

4. Descriptive table

Table 1. Mean Sleep Hours among Participants Who Aged 18-80 without Trouble Sleeping. (N=4,014)

Variables	*N(%)*	*Sleep Hours (Mean, SD)*
Age (Mean, SD)	45.52 (16.99)	
< 31	972 (24.22%)	7.12 (1.26)
31 - 44	978 (24.36%)	6.92 (1.20)
44 - 58	1045 (26.03%)	6.89 (1.20)
≥ 58	1019 (25.39%)	7.28 (1.28)
Gender		
Male	2262 (56.35%)	6.95 (1.21)
Female	1752 (43.65%)	7.19 (1.28)
PhysActive		
Yes	2252 (56.10%)	7.10 (1.15)
No	1762 (43.90%)	7.00 (1.35)
AlcoholYear (Median, IQR)	24 (3,104)	
≤ 24	2104 (52.42%)	6.99 (1.33)
> 24	1910 (47.58%)	7.13 (1.13)
Depressed		
Sometimes	642 (15.99%)	6.82 (1.41)
Never	3372 (84.01%)	7.10 (1.21)
Poverty (Mean, SD)	3.09 (1.62)	
≤ 1	530 (13.20)	7.08 (1.41)
> 1	3484 (86.80%)	7.05 (1.22)

Figure 2: Descriptive table