# STA490_Winter_Data_Analysis

## 1.Goal of Data Analysis

In the EDA phase, I found that the exercises, alongside with confounders such as number of stressors, average hours of sleep and BMI, have certain degrees of effects on the mental health status of subjects/students. In this data analysis, I want to further investigate whether or not the effects of these variables are statistically significant on the topic of investigation.

## 2.How you have chosen your model(s)

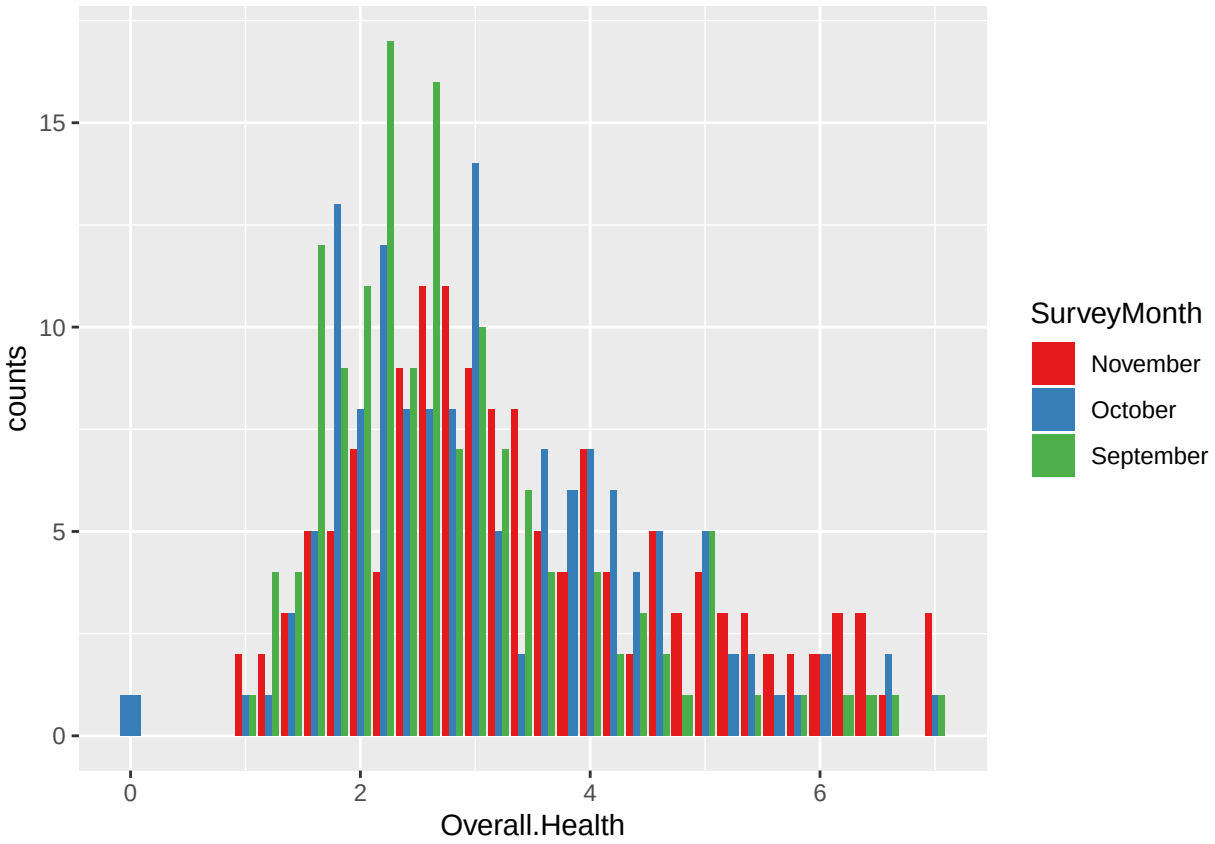
### 2.1.Response variable: Mental Health Indicator (Likert Scale 1-7)

I have created a mental health indicator using question 15-19 of the questionnaire, which assess the subject's concentration, energy level, feeling and sleep quality. **On a likert scale of 1-7, the lower the indicator the subjects are at, the better mental state they are at.**

```
data$Overall.Health = rowMeans(data[,c("Health1","Health2","Health3","Health4","Health5")])

data1 <- data %>%
  group_by(Overall.Health,SurveyMonth) %>%
  summarise(counts = n())

ggplot(data1, aes(x=Overall.Health, y= counts,fill = SurveyMonth)) +
  geom_bar(stat="identity", position = "dodge") +
  scale_fill_brewer(palette = "Set1")
```




## 2.2.Predictor variables: Amount of Exercises

According to the WHO guideline, adults aged 18–64 should do at least 150 minutes of moderate-intensity aerobic physical activity throughout the week or do at least 75 minutes of vigorous-intensity aerobic physical activity throughout the week or an equivalent combination of moderate- and vigorous-intensity activity. Hence, we will include those who have met these requirements into the treatment group.

```
data <- data %>%
  mutate(Treatment = ifelse(Exercise2==150 | Exercise3==75,1,0| (Exercise2/2 + Exercise3)>=75))

data$Treatment[2] = 0
data$Treatment[6] = 0
```

## 2.3.Confounders

In the EDA phase, I have found that the following confounders seem to have a significant effect on the mental health status of the subjects. Hence, they are included in our model.

```
data %>% group_by(BMI) %>%
  summarize(means = mean(Overall.Health))
```

```
## # A tibble: 5 x 2
##   BMI       means
##   <fct>     <dbl>
## 1 <18.5      3.26
## 2 >29.9      5
```

```
## 3 0          4.37
## 4 18.5-24.9  2.94
## 5 25-29.9    3.58
```

```
data %>% group_by(AvgHoursOfSleep) %>%
  summarize(means = mean(Overall.Health))
```

```
## # A tibble: 8 x 2
##   AvgHoursOfSleep means
##   <fct>           <dbl>
## 1 <4               4.89
## 2 >9               3.43
## 3 0                2.33
## 4 5                3.79
## 5 6                3.31
## 6 7                2.96
## 7 8                2.65
## 8 9                2.74
```

```
data %>% group_by(NumStressors) %>%
  summarize(means = mean(Overall.Health))
```

```
## # A tibble: 6 x 2
##   NumStressors means
##   <fct>        <dbl>
## 1 >4            4.24
## 2 0             2.29
## 3 1             2.45
## 4 2             3.04
## 5 3             3.13
## 6 4             3.27
```

## 2.4.Random Effect

In this experiment, the 420 observations come from 140 subjects who have completed 3 surveys in the span of 3 months Therefore, to prevent pseudoreplication, I wanted to treat the subject ID as a random effect. Beside subject ID, I also found the difference in month of survey also contributes significant variability(54.05393) in the mental health indicator, hence it is also included as a random effect.

## 2.5.Final Model

```
# final model
mod0 <- lm(Overall.Health~Treatment,data=data)

mod1 <- nlme::lme(Overall.Health~Treatment+
                    as.factor(BMI)+
                    as.factor(AvgHoursOfSleep)+
                    as.factor(NumStressors)+
                    as.factor(SurveyMonth),
                  random=list(~1|study.ID),
                  data=data)
```

### 2.4.Random Effect

In this experiment, the 420 observations come from 140 subjects who have completed 3 surveys in the span of 3 months Therefore, to prevent pseudoreplication, I wanted to treat the subject ID as a random effect. Beside subject ID, I also found the difference in month of survey also contributes significant variability(54.05%) in the mental health indicator, hence it is also included as a random effect.

```
knitr::kable(VarCorr(mod1))
```

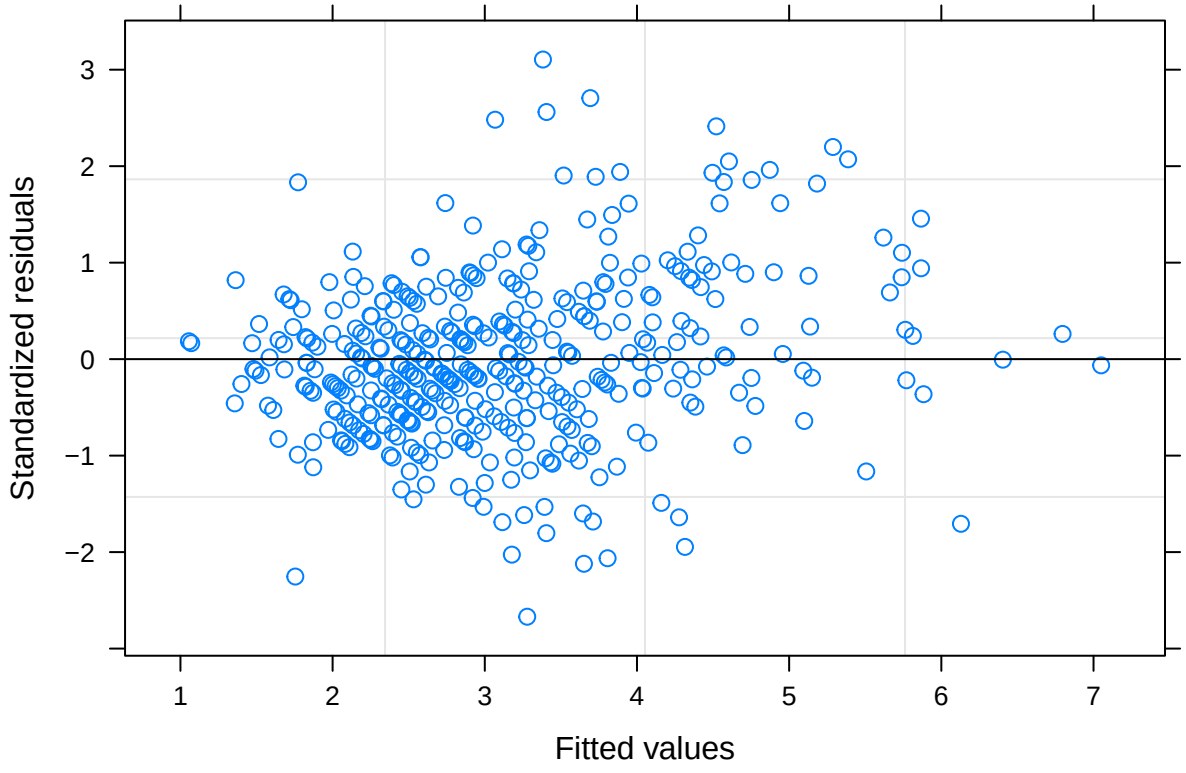
| | Variance | StdDev |
|---|---|---|
| (Intercept) | 0.6685579 | 0.8176539 |
| Residual | 0.6063341 | 0.7786746 |

## 3.Verify model assumptions and fits

There are a few model assumptions I want to check: 1. Homogeneity of variance 2. Normality of error term

### 3.1 Homogeneity of variance

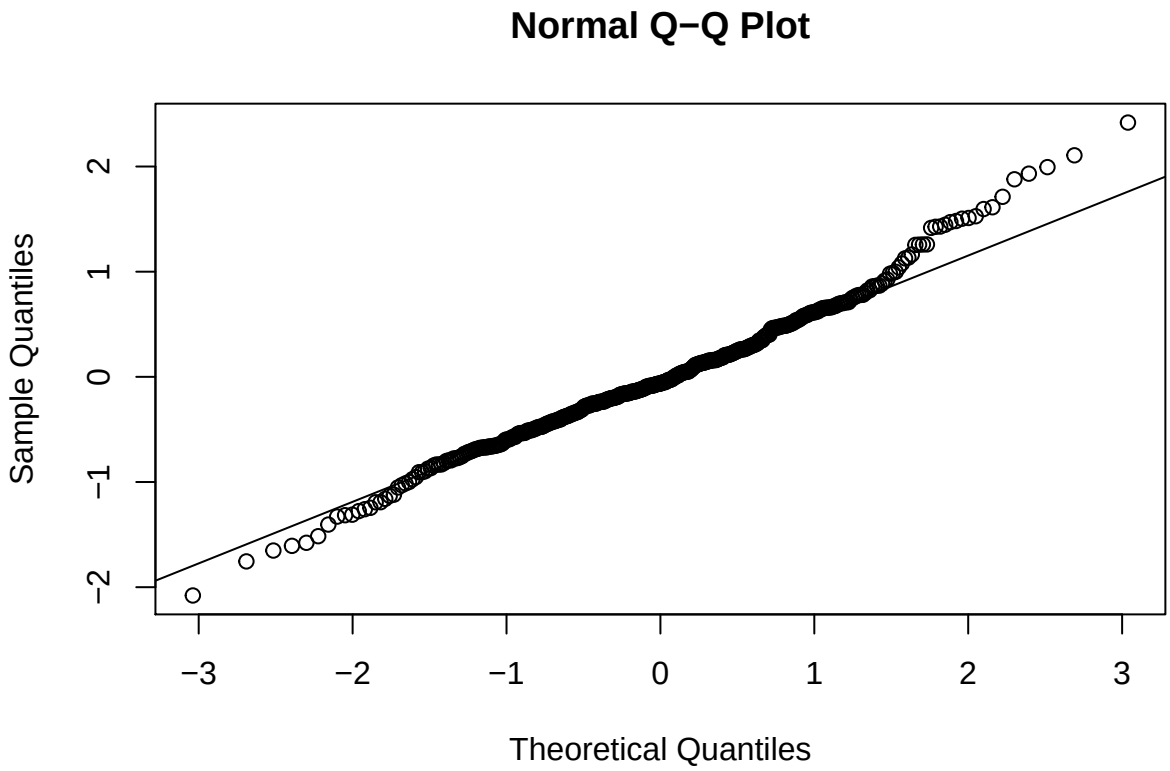
```
plot(mod1)
```




This residual plot does not indicate any deviations from a linear form. It also shows relatively constant variance across the fitted range. The slight reduction in apparent variance on the right f the graph are likely a result of there being fewer observations in these predicted areas.

### 3.2 Normality of error terms

```
qqnorm(residuals(mod1))
qqline(residuals(mod1))
```




The Q-Q Normal Plots do not raise any significant concern with normality of the weighted residuals.

## 4.What you have learned about the answer to the research question

```
# fixed effect outputs
table0 = cbind(summary(mod0)$coef[,-3],confint(mod0))
knitr::kable(table0,digits=3)
```

| | Estimate | Std. Error | Pr(>\|t\|) | 2.5 % | 97.5 % |
|---|---|---|---|---|---|
| (Intercept) | 3.150 | 0.080 | 0.00 | 2.994 | 3.307 |
| Treatment | -0.061 | 0.138 | 0.66 | -0.333 | 0.211 |

| | Estimate | Std. Error | Pr(>\|t\|) | 2.5 % | 97.5 % |
|---|---|---|---|---|---|

```
table1 = cbind(summary(mod1)$tTable[,-c(3,4)],intervals(mod1)$fixed[,-2])

knitr::kable(table1,digits=3)
```

| | Value | Std.Error | p-value | lower | upper |
|---|---|---|---|---|---|
| (Intercept) | 5.168 | 0.339 | 0.000 | 4.501 | 5.835 |
| Treatment | -0.046 | 0.116 | 0.692 | -0.276 | 0.183 |
| as.factor(BMI)>29.9 | 0.613 | 0.387 | 0.115 | -0.150 | 1.375 |
| as.factor(BMI)0 | 0.478 | 0.578 | 0.409 | -0.660 | 1.616 |
| as.factor(BMI)18.5-24.9 | -0.128 | 0.216 | 0.553 | -0.554 | 0.297 |
| as.factor(BMI)25-29.9 | 0.197 | 0.284 | 0.489 | -0.362 | 0.755 |
| as.factor(AvgHoursOfSleep)>9 | -0.597 | 0.464 | 0.199 | -1.510 | 0.316 |
| as.factor(AvgHoursOfSleep)0 | -1.739 | 0.576 | 0.003 | -2.874 | -0.604 |
| as.factor(AvgHoursOfSleep)5 | -0.741 | 0.287 | 0.010 | -1.306 | -0.176 |
| as.factor(AvgHoursOfSleep)6 | -0.963 | 0.278 | 0.001 | -1.510 | -0.416 |
| as.factor(AvgHoursOfSleep)7 | -1.317 | 0.280 | 0.000 | -1.867 | -0.766 |
| as.factor(AvgHoursOfSleep)8 | -1.351 | 0.298 | 0.000 | -1.938 | -0.764 |
| as.factor(AvgHoursOfSleep)9 | -1.401 | 0.353 | 0.000 | -2.095 | -0.707 |
| as.factor(NumStressors)0 | -1.001 | 0.211 | 0.000 | -1.417 | -0.585 |
| as.factor(NumStressors)1 | -1.044 | 0.167 | 0.000 | -1.373 | -0.714 |
| as.factor(NumStressors)2 | -0.642 | 0.156 | 0.000 | -0.949 | -0.335 |
| as.factor(NumStressors)3 | -0.722 | 0.156 | 0.000 | -1.028 | -0.415 |
| as.factor(NumStressors)4 | -0.587 | 0.194 | 0.003 | -0.970 | -0.204 |
| as.factor(SurveyMonth)October | -0.253 | 0.097 | 0.009 | -0.443 | -0.063 |
| as.factor(SurveyMonth)September | -0.379 | 0.101 | 0.000 | -0.578 | -0.179 |

```
knitr::kable(anova(mod1),digits=3)
```

| | numDF | denDF | F-value | p-value |
|---|---|---|---|---|
| (Intercept) | 1 | 261 | 1575.301 | 0.000 |
| Treatment | 1 | 261 | 0.476 | 0.491 |
| as.factor(BMI) | 4 | 261 | 5.977 | 0.000 |
| as.factor(AvgHoursOfSleep) | 7 | 261 | 8.916 | 0.000 |
| as.factor(NumStressors) | 5 | 261 | 11.915 | 0.000 |
| as.factor(SurveyMonth) | 2 | 261 | 7.507 | 0.001 |

```
knitr::kable(table1[c(5,6,3),],digits=3)
```

| | Value | Std.Error | p-value | lower | upper |
|---|---|---|---|---|---|
| as.factor(BMI)18.5-24.9 | -0.128 | 0.216 | 0.553 | -0.554 | 0.297 |
| as.factor(BMI)25-29.9 | 0.197 | 0.284 | 0.489 | -0.362 | 0.755 |
| as.factor(BMI)>29.9 | 0.613 | 0.387 | 0.115 | -0.150 | 1.375 |

```
knitr::kable(table1[c(9:13,7),],digits=3)
```

| | Value | Std.Error | p-value | lower | upper |
|---|---|---|---|---|---|
| as.factor(AvgHoursOfSleep)5 | -0.741 | 0.287 | 0.010 | -1.306 | -0.176 |
| as.factor(AvgHoursOfSleep)6 | -0.963 | 0.278 | 0.001 | -1.510 | -0.416 |
| as.factor(AvgHoursOfSleep)7 | -1.317 | 0.280 | 0.000 | -1.867 | -0.766 |
| as.factor(AvgHoursOfSleep)8 | -1.351 | 0.298 | 0.000 | -1.938 | -0.764 |
| as.factor(AvgHoursOfSleep)9 | -1.401 | 0.353 | 0.000 | -2.095 | -0.707 |
| as.factor(AvgHoursOfSleep)>9 | -0.597 | 0.464 | 0.199 | -1.510 | 0.316 |

```
knitr::kable(table1[c(15:18),],digits=3)
```

| | Value | Std.Error | p-value | lower | upper |
|---|---|---|---|---|---|
| as.factor(NumStressors)1 | -1.044 | 0.167 | 0.000 | -1.373 | -0.714 |
| as.factor(NumStressors)2 | -0.642 | 0.156 | 0.000 | -0.949 | -0.335 |
| as.factor(NumStressors)3 | -0.722 | 0.156 | 0.000 | -1.028 | -0.415 |
| as.factor(NumStressors)4 | -0.587 | 0.194 | 0.003 | -0.970 | -0.204 |

```
knitr::kable(table1[c(19,20),],digits=3)
```

| | Value | Std.Error | p-value | lower | upper |
|---|---|---|---|---|---|
| as.factor(SurveyMonth)October | -0.253 | 0.097 | 0.009 | -0.443 | -0.063 |
| as.factor(SurveyMonth)September | -0.379 | 0.101 | 0.000 | -0.578 | -0.179 |

Looking at the fixed effects due to treatment, the estimate is -0.0765 (p-value=0.516),indicating that there is no significant difference in the mental health status between the those who have exercised the amount based on WHO recommendation and those who have not.

However, if we look at our confounder variables, they all have indicated a certain degree of statistical significance, especially for the number of stressor. Looking at the estimates for the number of stressors, there seems to be a distinct trend: with 5 stressors at the baseline estimate, the less stressors the subjects have, the better their mental health status (lower indicators). The same trend happens with the average hours of sleep, the more hours of sleep the subjects have, the better their mental health status. On the other side, the trend in BMI is not as distinct, but from the estimate for those with BMI >29.9, they generally have worst mental health status(p-value=0.0532) than the baseline BMI.

## 5.Potential limitations in your work that you identified during your analysis.

There are a few limitations to the study:

- By fitting subject ID and survey of monthas as random intercepts, although differences among individuals can be obtained, this also uses a degree of freedom for these variables, severely limiting the power of the model.
- Since our response variable is intially a ranking data("Strongly Agree" to "Strongly Disagree"), we have to be careful in translating it into numerical values, as we have to make sure the new number scale is reflective of the true responses. In these ranking data, we also encounter missing values. These missing

values are also difficult to deal with, as sometimes missing values could be due to errors in the data collection process, or subjects miss out on some questions intentionally.