Final Project

Team B - Yuval Sabag, Shani Bahat

1. Introduction

Our data analyze project is based on Dr. Noga Cohen research – "Psychological experiment about social emotion regulation". The general research question we chose based on the research is-

How taking emotion regulation training affected the participants emotional regulation indexes With emphasis on rumination - the tendency to think negatively (RRS) and reduce depression (CES.D).

The specific question we have chosen to research and build a prediction model for is-

How taking emotion regulation training affected the participants RRS indexes in the long term based on the initial RRS and CES.D indexes and the gender

In the project, we will examine the participants emotional regulation indexes in three periods of time - before the training, at the end the training and two months after the training. Mainly, we will examine the effect regarding to participant gender.

This data analyze helps to understand whether it is possible to regulate emotions with professional guidance in a long-term view. We have chosen to focus on two main indexes:

- **RRS** - Rumination - the tendency to think circularly and repetitively about past events.
  According to studies, rumination thinking leads to continuous negative thoughts, loss of perspective and focusing on negative aspects. Thus, instead of producing solutions to ones problems, rumination inhibits or even prevents them.

- **CES.D** - depression experienced by the participant.
  Depression is a mood disorder that causes a persistent feeling of sadness and loss of interest.

Rumination is a predictor of depression and one of the most problematic cognitive symptoms associated with it. Previous research has demonstrated significant relationships between rumination and depression. Based on these we will examine whether it is possible to change a person's negative view of events he has experienced.

The data analyze results will help us to understand whether people can be trained to regulate their own emotions – a conclusion that can improve the quality of life for many, and a subject that were set by the UN as one of the "Sustainable Development Goals" (part of Good health and well-being). Emotional regulation is a topic that has not been studied much before and therefore there are only few studies on the subject.

Our approach is to use the questionnaires provided to participants and the data provided to us. Based on that we will analyse the emotional state of the participants on three time periods, the data analysis will help us determine the effectiveness of the training and the ability to regulate emotions.

2. Data Overview

The data we will use contains the participants answers to an online survey that was taken on three different time periods of the trial:

Pre training (there were two different tables for self/provider- we combined them and used them as one)
Post training
Follow up - 2 months after the trail.

Each online survey was identical therefore each table has identical columns.

The data contains information about emotion regulation skills and their maintenance over time. In our analysis we will focus on two main indicators:

- **RRS** - rumination (tendency to think negatively about past events – bad strategy)
  There are 22 columns in each table that refer to this index. In order to analyze this data, we use the columns amount.
- **CES.D** - depression
  There are 20 columns in the table that refer to this index.
  In order to analyze this data, we use the columns amount.

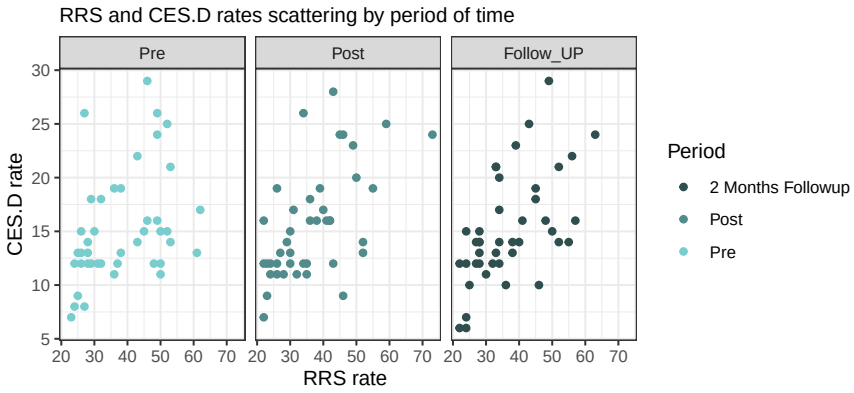
In addition, all tables include individual characteristics of the participants, that will allow us to analyze the effect of the experiment on different groups in the population: gender, age, SES - socioeconomic status, ethnicity, education.

for our analyze we combined data from the 3 given data sets into one table that contains personal characteristics of each participant as well as the sum of RRS and CES.D columns for each period of time (one row = one participant). After cleaning the data and removing outliers our data contain 42 reliable observation.

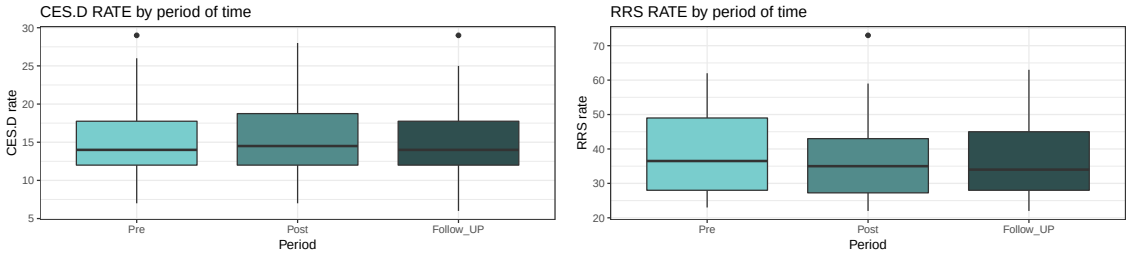
3. Methods and results

We started with plotting the initial data in order to infer some conclusion about the parameters scattering and to identify connections.

This plot helps us to infer that there is a positive correlation between RRS and CES.D.



From the below analysis we notice an improvement in the RRS index compared to the CES.D index which have a minor improvement.



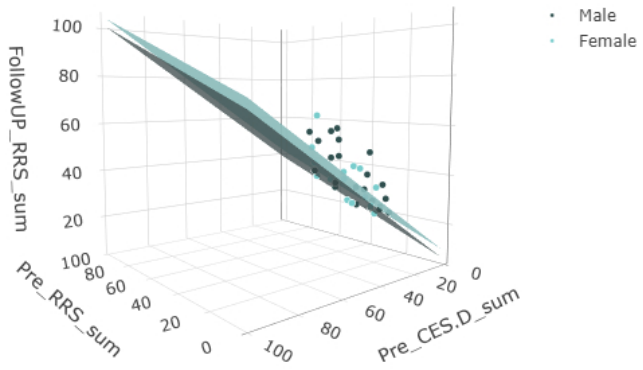
Based on the initial data analysis we chose to construct our model as a multiple linear regression that will predict the effect of the experiment on the long-term RRS index. The two main variables on which the model is based on are the initial RRS and CES.D rates.

In order to test our model we have split it into train and test with 0.9 proportion. That will allow us to check the model's accuracy.

We tried to add several personal characteristics to the model and found out that the adjusted R squared index is the best when combined the initial RRS and CES.D measurements with the gender (factorial variable) .

Eventually we got the following model (with adjusted R squared = 0.46):

Long term RRS = 9.0327 + 0.6293 x initial RRS + 0.3277 x initial CES.D - 3.3736 x Gender (Male = 1, Female = 0)



To test the model we used the test data and got the following results:

RMSE = 2.36558 **R squaerd** = 0.9661014

By those measurements we can tell that the model is relatively a good fit with quite high variance.

Unlike our hypotheses, the model predicts that the treatment is not always useful and doesn't have as much influence as we expected.

The RRS rate can be higher or lower then the initial RRS rate based on the participant initial RRS and CES.D measurements as well as his gender. The combination of those three will help us to predict whether the emotional regulation will help the participant to reduce his tendency to rumination or should he seek for other treatments. Additionally we noticed that the treatment affects men more than women.

***our full data analysis can be found in the attached code.

4. Limitations and Future work

One of the major limits that our project faces is that it relies on relatively small dataset. Therefore, the results are not as accurate as we wish them to be and the conclusions are not reliable enough to reflect and generalize to other populations. Additionally, mental states and emotions are hard to predict and are often given a different interpretation by different people. Therefore we expect to get general tendency and not accurate results.

If we had additional three month we would conduct the experiment once more on a larger population. This would allow us to base our model on more observations and to improve the model's accuracy. In addition we would have modify the questions so they would be less subject to personal interpretation, and focus only on the questions that are relevant to our chosen indexes (RRS and CES.D).

Appendix

Combined final data

```
## Rows: 42
## Columns: 10
## Rowwise:
## $ WorkerID               <chr> "A149ROBL26JWPJ", "A36470UBRH28GO", "A1ZT30BGR3~
## $ group                  <chr> "self", "self", "self", "self", "self", "self",~
## $ Gender                 <fct> Male, Female, Female, Female, Female, Female, M~
## $ SES_Degree             <dbl> 3, 3, 2, 1, 1, 3, 2, 3, 1, 3, 3, 4, 4, 3, 3, 3,~
## $ Pre_RRS_sum            <dbl> 49, 52, 25, 43, 36, 38, 29, 32, 49, 26, 61, 62,~
## $ Pre_CES.D_sum          <dbl> 24, 25, 13, 14, 19, 13, 18, 18, 16, 13, 13, 17,~
## $ PostTraining_RRS_sum   <dbl> 59, 46, 30, 28, 22, 42, 31, 39, 45, 24, 36, 52,~
## $ PostTraining_CES.D_sum <dbl> 25, 24, 13, 11, 16, 16, 17, 19, 24, 12, 16, 13,~
## $ FollowUP_RRS_sum       <dbl> 52, 49, 28, 38, 27, 41, 34, 33, 56, 22, 34, 55,~
## $ FollowUP_CES.D_sum     <dbl> 21, 29, 13, 13, 14, 16, 14, 13, 22, 12, 20, 14,~
```

Summary Statistics

```
## # A tibble: 3 x 3
##   Period    mean_RRS mean_CES.D
##   <chr>        <dbl>      <dbl>
## 1 Follow_UP     36.9       15.1
## 2 Post          36.7       15.6
## 3 Pre           38.2       15.2
```

Linear regression detalis

```
##
## Call:
## lm(formula = FollowUP_RRS_sum ~ Pre_RRS_sum + Pre_CES.D_sum +
##     Gender, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4301  -4.2813  -0.8275   5.5264  17.1060
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     9.0327     5.2972   1.705   0.0976 .
## Pre_RRS_sum     0.6293     0.1363   4.618 5.66e-05 ***
## Pre_CES.D_sum   0.3277     0.2890   1.134   0.2649
## GenderFemale   -3.3736     2.8694  -1.176   0.2481
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.257 on 33 degrees of freedom
## Multiple R-squared:  0.5037, Adjusted R-squared:  0.4586
## F-statistic: 11.16 on 3 and 33 DF,  p-value: 3.262e-05
```

Predict on the testing dataset

```
## # A tibble: 5 x 5
##   .pred FollowUP_RRS_sum Pre_RRS_sum Pre_CES.D_sum Gender
##   <dbl>            <dbl>       <dbl>         <dbl> <fct>
## 1  25.7               28          25            13 Female
## 2  33.2               34          29            18 Male
## 3  43.3               46          52            15 Female
## 4  30.3               28          26            15 Male
## 5  47.0               50          53            14 Male
```