MSCI 718 2023W Assignment 2 - Partial Correlation and Bootstrapping

Data Summary

The dataset refers to empathy, mental health, and burnout of medical students in Switzerland. It has 20 variables and 886 observations. The report includes important demographic information as well as self-reported data and results from psychological tests to provides a comprehensive picture of the mental states of students in the medical field. The focus of this report is mainly on:

- **year**: (categorical, ordinal, int) Year of study of participants, year of study 1-3 are **B.MED** students (523 observations) with **age** (min: 21, max: 44, mean: 24.38, median: 24) and that 4-6 years are **M.MED** students (363 observations) with **age** (min: 17, max: 49, mean: 21, median: 21)
- **cesd**: (interval, int) Center for Epidemiologic Studies Depression scale (min: 0, max: 56, mean: 19.98, median: 18 for **B.MED**) and (min: 0, max: 54, mean: 15.26, median: 13 for **M.MED**)
- **stai_t**: (interval, int) State-Trait Anxiety Inventory scale of the participant (min: 20, max: 77, mean: 42.9, median: 43 for **B.MED**) and (min: 20, max: 69, mean: 40.78, median: 41 for **M.MED**)
- **health**: (categorical, ordinal, int) Self-reported health status of the participant (1 -> Very dissatisfied to 5 -> Very satisfied) (min: 1, max: 5, mean: 3.713, median: 4 for **B.MED**) and (min: 1, max: 5, mean: 3.871, median: 4 for **M.MED**)

This report is aimed at showing **How CESD is correlated to STAI_T in participants studying B.MED and M.MED, as well as when "controlled by HEALTH"**.

Planning

In order to select the correct correlation test, we need to consider and check a few assumptions. The two columns we are working (**cesd** and **stai**) on are both of "int" type and therefore of "interval" type. Now, let's test the normality of our columns. The Shapiro-Wilk normality test for (cesd) and (stai) shows that we have enough evidence to reject the null hypothesis (that the column has a normal distribution) at the 95 percent significance level, and thus conclude that they are not normal (p\~=0) (Please see Appendix 3 and 4).

QQ plots are calculated to further verify and ensure this output (Please see Appendix 5, 6, 7 and 8 for plots). It seems reasonable to interpret from these plots that none of these columns follow to the normal distribution.

Therefore, we must either change our test or our data. Now that these columns can be attributed to transformations like log, sqrt, or reciprocal, the Shapiro-Wilk results do not improve in favour of normality (details of the resulting transformations can be found in Appendix 9 and 10).

Data correlation boot strapping is performed because the data is not normal even after data conversion. The results are presented in Figures 1 and 2.

```
Call:
boot(data = B.MED_data, statistic = boot_func, R = 1000)
Bootstrap Statistics :
     original        bias     std. error
t1* 0.7256473 -0.0006518856  0.02512638
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot_B.MED, conf = 0.95, type = "perc")

Intervals :
Level     Percentile
95%   ( 0.6729,  0.7716 )
Calculations and Intervals on Original Scale
```

Figure 1: Bootstrapping result for B.MED

```
Call:
boot(data = M.MED_data, statistic = boot_func, R = 1000)
Bootstrap Statistics :
     original      bias     std. error
t1* 0.6803603 0.000758586  0.02931579
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot_M.MED, conf = 0.95, type = "perc")

Intervals :
Level     Percentile
95%   ( 0.6269,  0.7400 )
Calculations and Intervals on Original Scale
```

Figure 2: Bootstrapping result for M.MED

Analysis

Evidently, non-parametric tests must be used to test the partial correlation of these two columns (Since the number of rows in our newly modified dataset is well beyond 30 (n\>30), it is reasonable to assume that this dataset follows a normal distribution based on the central theorem limit). As a result, **Spearman rank-order correlation** is used**,** and the results are described in Figure 3, 4 and 5. Both tests unanimously agree that there is a **strong (rho > 0.5) positive correlation** between two columns. In other words, we have enough evidence to reject the null hypothesis of these two variables (CESD and STAI_T) correlation tests.

```
        Spearman's rank correlation rho
data:  B.MED_health$cesd and B.MED_health$stai_t
S = 6772531, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7159474
```

Figure 3: Correlation of CESD and STAI_T for B.MED

```
        Spearman's rank correlation rho
data:  M.MED_health$cesd and M.MED_health$stai_t
S = 2286188, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7132215
```

Figure 4: Correlation of CESD and STAI_T for M.MED

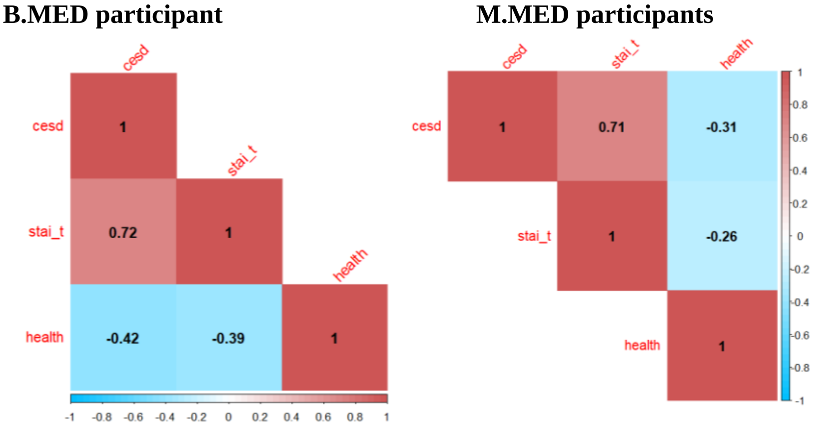


Figure 5: Correlation of CESD, STAI_T and HEALTH by Spearman method

From Figure 5, it is evident that there is correlation between HEALTH, CESD and STAI_T. Furthermore, **partial correlation with Spearman** is used for two columns, CESD and STAI T, across B.MED and M.MED participants, while controlling for the third column, HEALTH, and the resulting matrix and plots are shown in Table 1 and Figure 6 and 7.

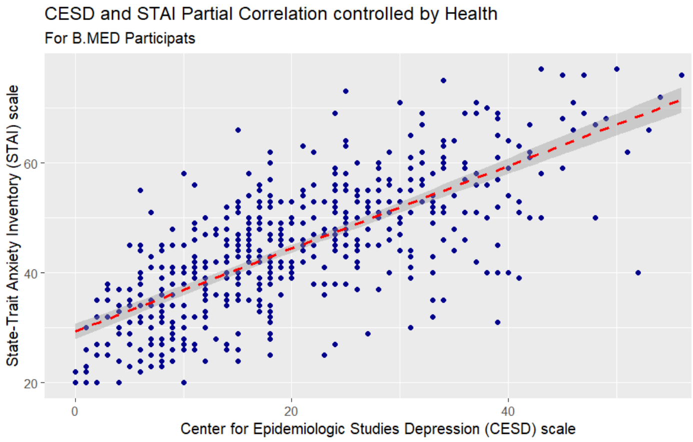


Figure 6: Partial Correlation plot for B.MED

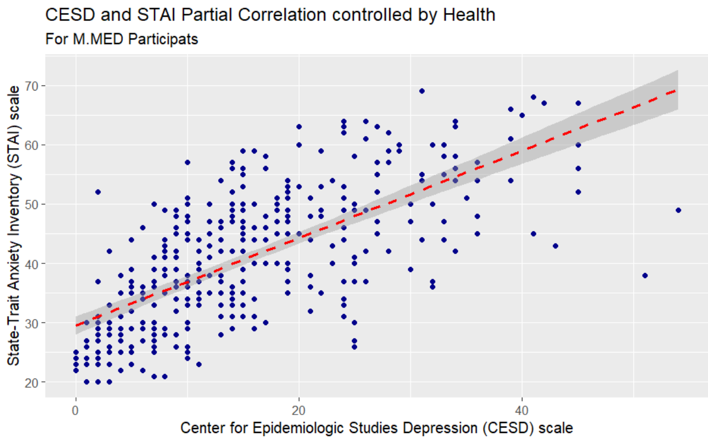


Figure 7: Partial Correlation plot for M.MED

Participants		Correlation Estimate		P. value		Statistic		N
	Variable	CESD	STAI_T	CESD	STAI_T	CESD	STAI_T	
B.MED	CESD	1	0.6611648	0	6.709363e-67	0	20.096053	523
	STAI_T	0.6611648	1	6.709363e-67	0	20.096053	0	
	HEALTH	-0.2142702	-0.1440345	7.761609e-07	9.663801e-04	-5.002294	-3.31910	
M.MED	CESD	1	0.6884947	0	3.614085e-52	0	18.012303	363
	STAI_T	0.6884947	1	3.614085e-52	0	18.012303	0	
	HEALTH	-0.1867154	-0.058769	3.547043e-04	0.2647416	-3.606092	-1.116995	

Table 1: Partial Correlation of CESD and STAI controlled by Health

Conclusion

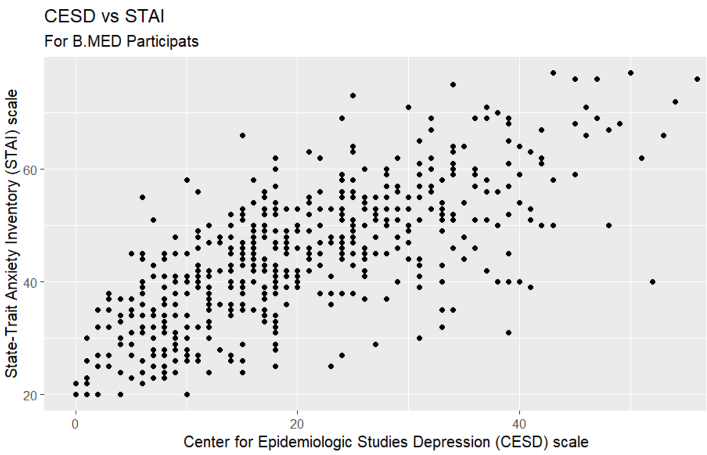
From the previous steps, it is clear that the **Spearman correlation between CESD and STAI_T** is (rs(523) = 0.7159, p < .001 for B.MED) and (rs(363) = **0.7132**, p < .001 for M.MED), indicating a **strong and positive correlation**.

However, using **bootstrapping**, rho is (rs(523) = **0.7256**, p < .001 for B.MED) and (rs(363) = **0.6804**, p < .001 for M.MED), indicating a strong correlation in both but a slight difference in correlation.

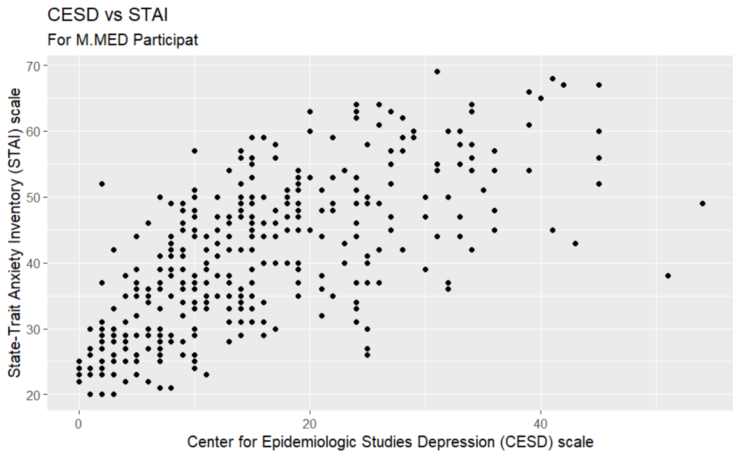
Furthermore, for B.MED, the partial correlation between **CESD and STAI_T controlled by HEALTH** is (rs(523) = **0.6612**, p < .001), whereas for M.MED, it is (rs(363) = **0.6885**, p < .001). Moreover, there is a **negative small-to-medium correlation** between HEALTH and CESD for both B.MED and M.MED is (rs(523) = **-0.2143**, p < .001 for B.MED) and rho is (rs(363) = **-0.1444**, p < .001 for M.MED),

While there is a negative small-to-medium correlation between HEALTH and STAI_T (rs(523) = **-0.1867**, p < .001) for both B.MED and while for M.MED (rs(363) = **-0.0588**, p = 0.245) with **p > 0.05** indicating that we do not have enough evidence to reject the null hypothesis in this case. It is important to note that there are also other factors besides HEALTH influencing the CESD and STAI_T in M.MED.

Appendix



Appendix 1: B.MED Data Visualization



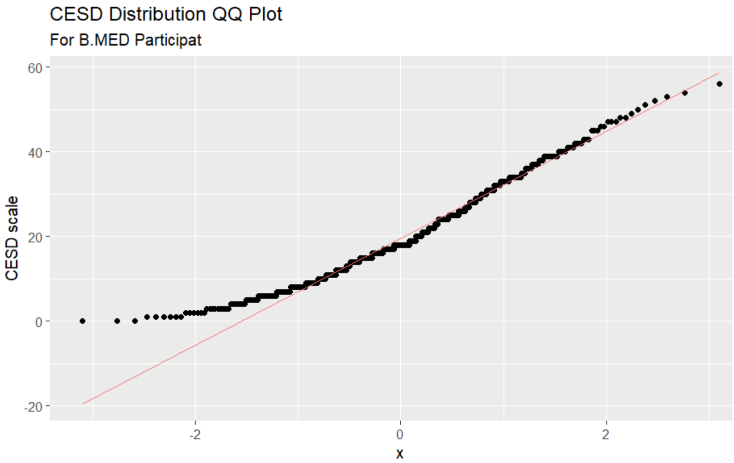
Appendix 2: M.MED Data Visualization

```
      cesd            stai_t           year           health           age    
 Min.   : 0.00   Min.   :20.00   Min.   :1.000   Min.   :1.000   Min.   :17  
 1st Qu.:11.00   1st Qu.:36.00   1st Qu.:1.000   1st Qu.:3.000   1st Qu.:19  
 Median :18.00   Median :44.00   Median :2.000   Median :4.000   Median :21  
 Mean   :19.98   Mean   :44.37   Mean   :1.805   Mean   :3.713   Mean   :21  
 3rd Qu.:28.00   3rd Qu.:52.00   3rd Qu.:3.000   3rd Qu.:4.000   3rd Qu.:22  
 Max.   :56.00   Max.   :77.00   Max.   :3.000   Max.   :5.000   Max.   :49  

        Shapiro-Wilk normality test

data:  B.MED_data$cesd
W = 0.96691, p-value = 1.795e-09


        Shapiro-Wilk normality test

data:  B.MED_data$stai_t
W = 0.98866, p-value = 0.0004463
```

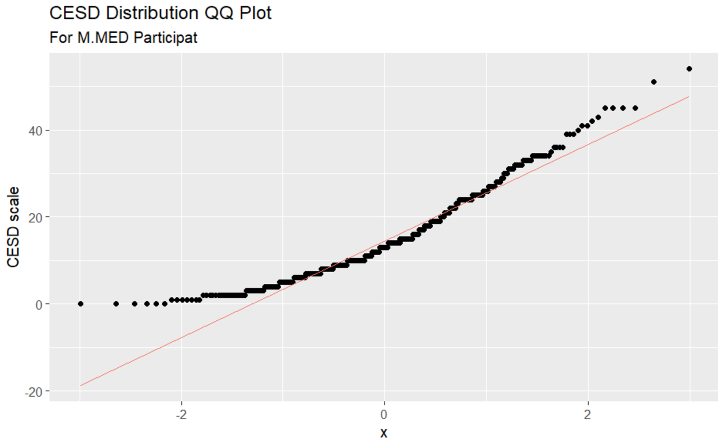
Appendix 3: B.MED Data summary and Shapiro-wilk test result for CESD and STAI_T

```
      cesd           stai_t           year           health           age       
 Min.   : 0.00   Min.   :20.00   Min.   :4.000   Min.   :1.000   Min.   :21.00  
 1st Qu.: 7.00   1st Qu.:31.00   1st Qu.:4.000   1st Qu.:3.000   1st Qu.:23.00  
 Median :13.00   Median :41.00   Median :5.000   Median :4.000   Median :24.00  
 Mean   :15.26   Mean   :40.78   Mean   :4.972   Mean   :3.871   Mean   :24.38  
 3rd Qu.:22.00   3rd Qu.:49.00   3rd Qu.:6.000   3rd Qu.:5.000   3rd Qu.:25.00  
 Max.   :54.00   Max.   :69.00   Max.   :6.000   Max.   :5.000   Max.   :44.00  

        Shapiro-Wilk normality test

data:  M.MED_data$cesd
W = 0.9347, p-value = 1.593e-11


        Shapiro-Wilk normality test

data:  M.MED_data$stai_t
W = 0.97621, p-value = 1.088e-05
```

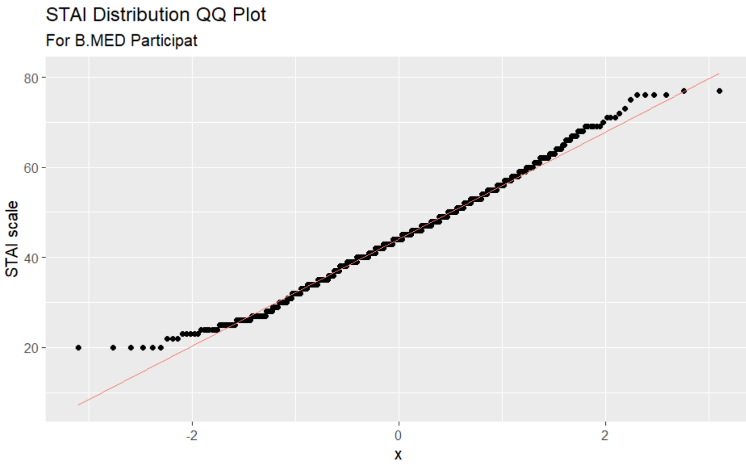
Appendix 4: M.MED Data summary and Shapiro-wilk test result for CESD and STAI_T



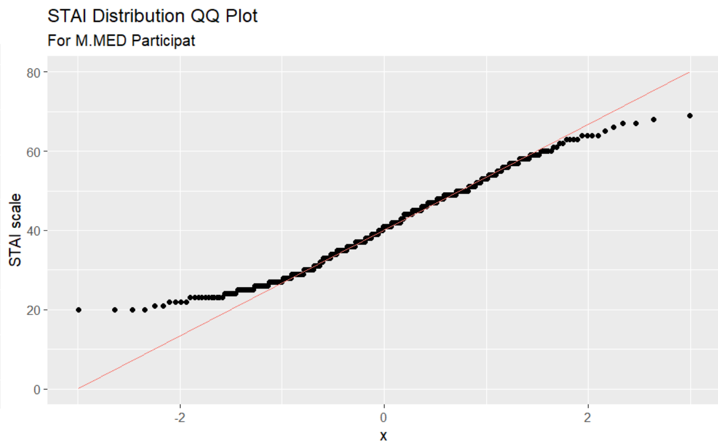
Appendix 5: CESD Distribution QQ plot for B.MED



Appendix 6: CESD Distribution QQ plot for M.MED



Appendix 7: STAI_T Distribution QQ plot for M.MED



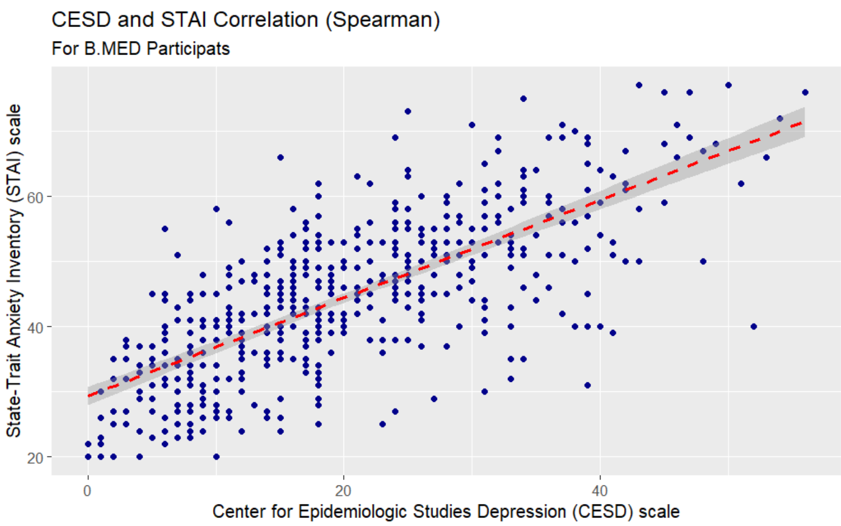
Appendix 8: STAI_T Distribution QQ plot for M.MED

```
        Shapiro-Wilk normality test

data:  cesd_sqr
W = 0.99264, p-value = 0.0113

        Shapiro-Wilk normality test

data:  cesd_reci
W = NaN, p-value = NA

        Shapiro-Wilk normality test

data:  cesd_log
W = NaN, p-value = NA

        Shapiro-Wilk normality test

data:  stai_sqr
W = 0.99196, p-value = 0.006333

        Shapiro-Wilk normality test

data:  stai_reci
W = 0.9144, p-value < 2.2e-16

        Shapiro-Wilk normality test

data:  stai_log
W = 0.98023, p-value = 1.572e-06
```

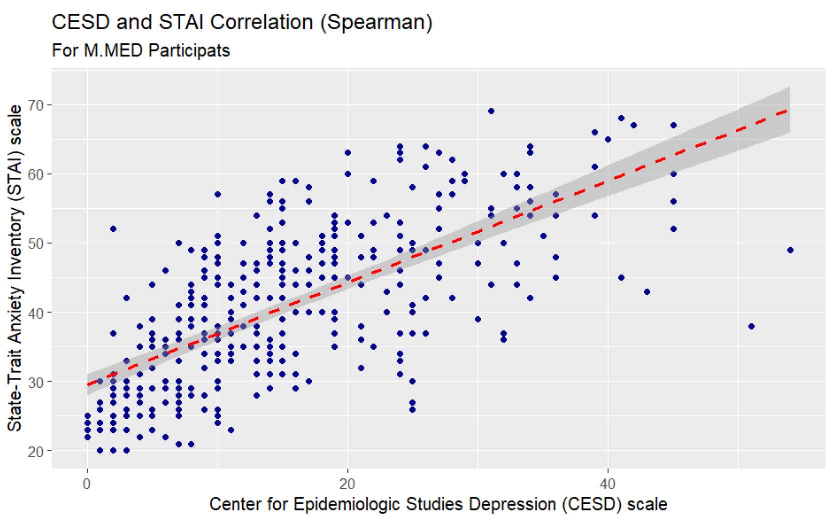
Appendix 9: B.MED CESD and STAI_T data conversion Shapiro-wilk normality test result

```
        Shapiro-Wilk normality test

data:  cesd_sqr
W = 0.99292, p-value = 0.08457


        Shapiro-Wilk normality test

data:  cesd_reci
W = NaN, p-value = NA


        Shapiro-Wilk normality test

data:  cesd_log
W = NaN, p-value = NA


        Shapiro-Wilk normality test

data:  stai_sqr
W = 0.97933, p-value = 4.548e-05


        Shapiro-Wilk normality test

data:  stai_reci
W = 0.93268, p-value = 9.729e-12


        Shapiro-Wilk normality test

data:  stai_log
W = 0.9729, p-value = 2.619e-06
```

Appendix 10: M.MED CESD and STAI_T data conversion Shapiro-wilk normality test result



Appendix 11: CESD and STAI_T Correlation for B.MED (Spearman method)



Appendix 12: CESD and STAI_T Correlation for M.MED (Spearman method)