Individual Assignment 2

Partial Correlation or Bootstrapping

Sina Bahrami

Part 1: Data Preprocessing

- The data frame is downloaded from [1] and includes data on mental health of medical students.
- The data frame has 886 rows and 20 columns.
- The variable names in the data frame are id, age, year, sex, glang, part, job, stud_h, health, psyt, jspe, qcae_cog, qcae_aff, amsp, erec_mean, cesd, stai_t, mbi_ex, mbi_cy, mbi_ea.
- There are 0 missing data.
- Also among the interval data, outliers are detected and removed. However, there were very few outliers as visualized by boxplot in the appendix.

Part 2: Planning

As an initial evalutation of association between data, the correlation matrix of interval data are calculated using Pearson's method and the matrix is plotted as heatmap in the appendix. The result suggests a good correlation between MBI-EX (a measure of emotional exhaustion) and CESD (a measure of depression) with a correlation coefficient of 0.585, so they are selected for the rest of the study. Stai_t and cesd were the other candidate pair but by the diffinition they are just different measures of depression, so naturally they should have high positive correlation, so it is not investigated here.

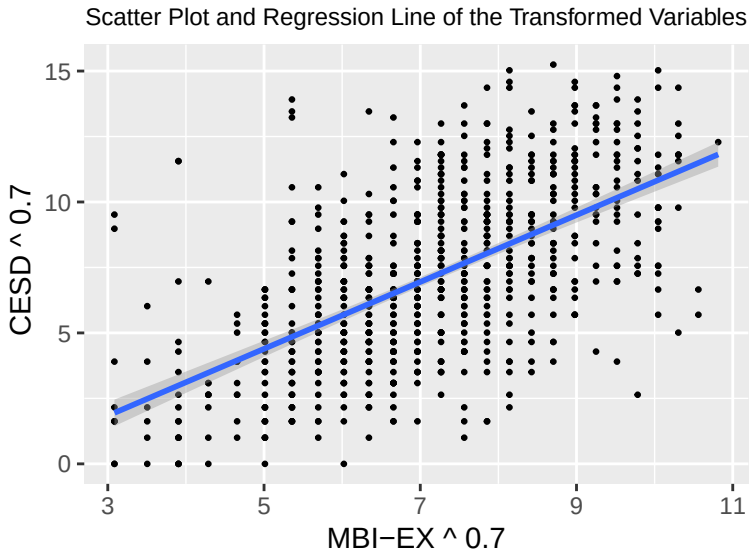
- qcae_aff: affective empathy score
- mbi_ex: MBI emotional exhaustion

Statistical summary of each data is provided in the appendix. It is worth mentioning that both variables are assumed interval variables.
Their correlation test requires test of normality. The qq-normal plot for both variables exhibit satisfactory normality. Histograms and QQ-nomral plot of both variables suggest that CESD is non-normal and is right-skewed, but MBI-EX seems almost normal. To ensure this, the result of normality test, Skewness.2SE and Kurtosis.2SE are observed. The results support the conclusion from the histogram plot. However, normality test rejects the normality of both variables.

- Despite the qq-normal plot, based on Shapiro-Wilk test, both MBI-EX and CESD are non-normal with p = 2.02e-05 and p = 1.03e-13 respectively i.e. $p* < 0.05$.
  However, the sample size is 810 (more than 30). This is an alternative condition for normality.
  To improve the situation for correlation test, a transformation is tried to make the data (particularly CESD) normal thus reducing error as much as possible.
- The transformation is aimed to improve skewness of CESD, but applying transformation only on CESD would make it difficult to study the linear relationship between CESD and MBI-EX.
- On the other hand, MBI-EX has already very little skewness, so applying the transformation on it will increase its skewness and reduce its normality. So, this trade-off needs to be considered in selection of the transformation function.
- The applied transformation is $X^{0.7}$. (Signs of variable values are all positive, so the transformation can be applied for the whole range without producing null values.)
- As shown in the QQ-normal plot and histogram in the appendix, the transformation has eliminated most of the skewness, but there is still some kurtosis for CESD. However, considering the satisfactory qq-normal and the good sample size (>30), the Pearson's correlation test is performed on the variable pair i.e. both assumptions of this test, which are interval data and normality (or sample size>30), are satisfied.

Part 3: Analysis

- The correlation coefficient is 0.601 with the 95% confidence interval of (0.556, 0.644).
- Because p-value is 7.63e-81 ($*p < 0.05$), the null hypothesis is rejected i.e. there is correlation between MBI emotional exhaustion and affective empathy at 5% level of significance.
- Instead of transforming data, we can change our method to Spearman, which does not required normality and interval data. The p-value of Spearman's method would be 4.16e-81, which also rejects Null hypothesis i.e. supports correlation.
- The confidence interval of Spearman's can be obtained from bootstrapping and be compared with the one from Pearson's.
- The confidence interval from basic bootstrapping of Spearman's method on non-transformed data is 0.554, 0.652. This CI is very close to the one from Pearson's method: 0.556, 0.644.



The association between the two variables is also visualized using scatter plot in combination with a linear model regression line. However, Regression models are very sensitive to homoscedasticity of variables. The p-value of homoscedasticity for the variables is 4.33e-75, so the null hypothesis is rejected and assumption of homoscedasticity is rejected, which means the linear regression model and the visualized CI is not valid.

Part 4: Conclusion

In this study two variables, MBI-EX and CESD, which represent work burnout in form of emotional exhaustion and depression respectively are studied for medical students. Data were almost clean with few outliers. Both variables were considered as interval data. To obtain the correlation between them, and carry out the suitable tests, the assumption were investigation with details provided in the appendix. The correlation test and correlation coefficient suggest that there is moderate to strong correlation between the variables. However, it is worth mentioning this only support association between the variables and not causation. Also using bootstrapping method, CI for Spearman's method was obtained which was very close to the one obtained from Pearson's test.

References

1. Sayadi, Fares. "Medical Student Mental Health." Kaggle, 25 Jan. 2023, https://www.kaggle.com/datasets/thedevastator/medical-student-mental-health/code.

Appendix

Codes

```
knitr::opts_chunk$set(echo = TRUE)
library(dplyr)
library(tidyr)
library(readr)
library(magrittr)
library(ggplot2)
library(knitr)
library(pastecs)
library(boot)
library(ggm)
library(car)
library(corrplot)

detect_outlier <- function(x) {

  Quantile1 <- quantile(x, probs=.25, na.rm = TRUE)
  Quantile3 <- quantile(x, probs=.75, na.rm = TRUE)
  IQR = Quantile3-Quantile1
  x > Quantile3 + (IQR*1.5) | x < Quantile1 - (IQR*1.5)
}

# Takes a data frame and a column and removes the outliers from the
# original data frame
remove_outlier <- function(dataframe,
                           columns=names(dataframe)) {
  # for loop to traverse in columns vector
  for (colmn in columns) {
    # remove observation if it satisfies outlier function
    dataframe <- dataframe[!detect_outlier(dataframe[[colmn]]), ]
  }
  return(dataframe)
}
df_main <- read_csv(file="Data Carrard et al. 2022 MedTeach.csv",
                    show_col_types=FALSE)

# Let's first take a look at the data
head(df_main)
```

```
## # A tibble: 6 x 20
##      id   age  year   sex glang  part   job stud_h health  psyt  jspe qcae_cog
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>  <dbl> <dbl> <dbl>    <dbl>
## 1     2    18     1     1   120     1     0     56      3     0    88       62
## 2     4    26     4     1     1     1     0     20      4     0   109       55
## 3     9    21     3     2     1     0     0     36      3     0   106       64
## 4    10    21     2     2     1     0     1     51      5     0   101       52
## 5    13    21     3     1     1     1     0     22      4     0   102       58
## 6    14    26     5     2     1     1     1     10      2     0   102       48
## # ... with 8 more variables: qcae_aff <dbl>, amsp <dbl>, erec_mean <dbl>,
## #   cesd <dbl>, stai_t <dbl>, mbi_ex <dbl>, mbi_cy <dbl>, mbi_ea <dbl>
```

```
# column number of the data of different types are separated
intervals <- c(2, 3, 8, 11:20)
nominals <- c(4, 5)
ordinals <- c(9)
```

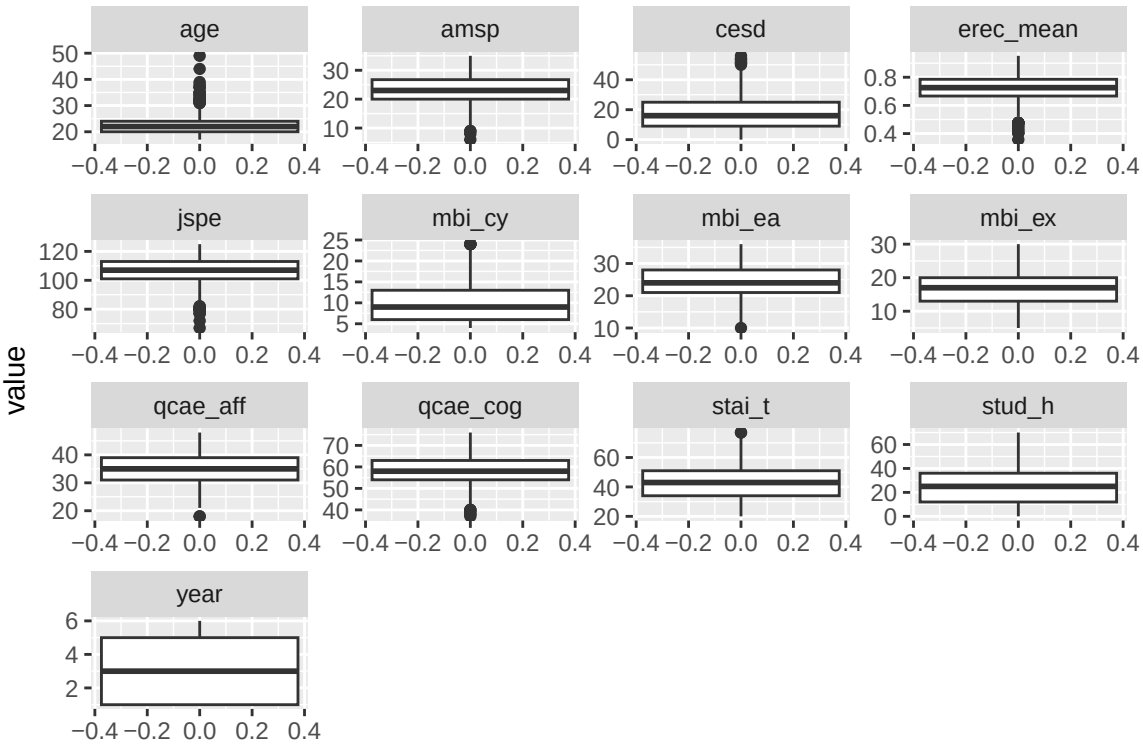
```
binaries <- c(6, 7, 10)

cnames <- names(df_main)

# We can visualized mbi_ex vs. qcae_aff to visually explore their relationship
sum(is.na.data.frame(df_main))
```

```
## [1] 0
```

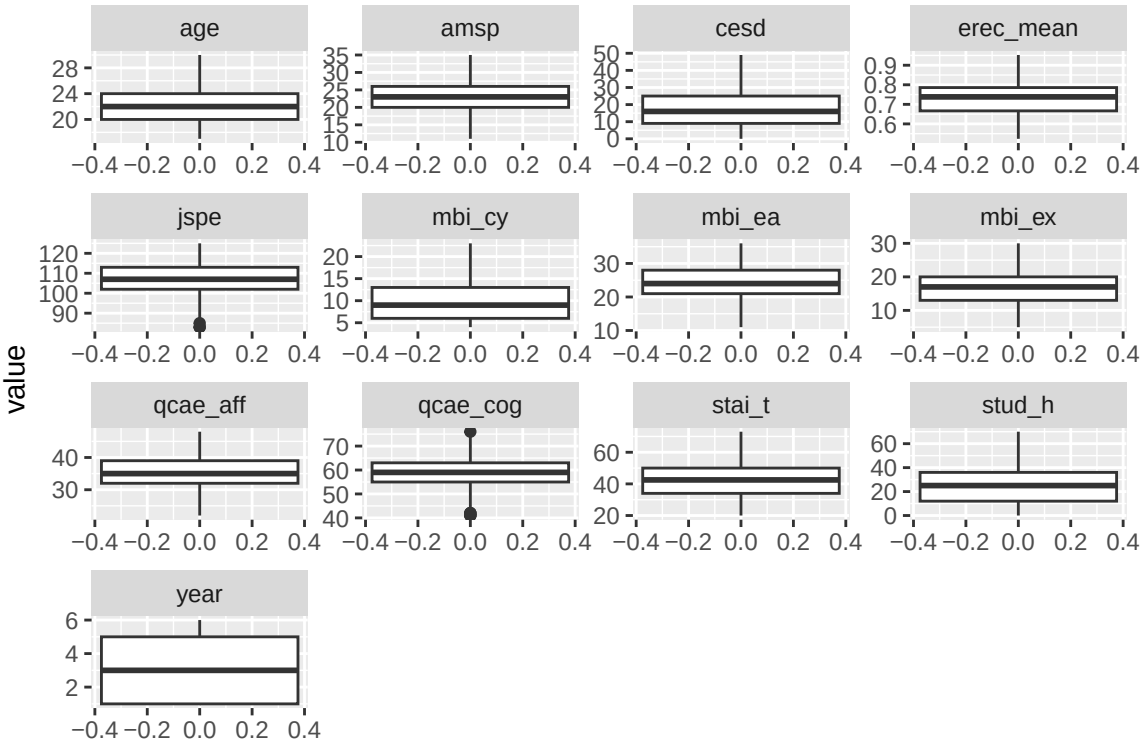
```
# Binary data are all 1 or 0
table(df_main[binaries])
```

```
## , , psyt = 0
##
##      job
## part   0   1
##    0 212  93
##    1 246 136
##
## , , psyt = 1
##
##      job
## part   0   1
##    0  50  32
##    1  69  48
```

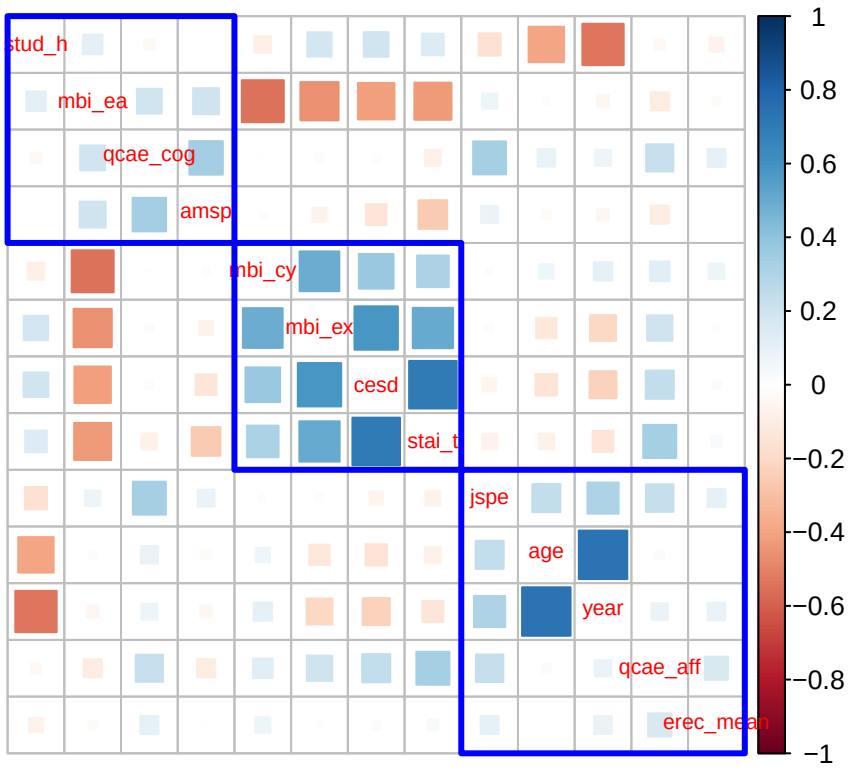
```
# Plotting outliers for interval data
df_main %>%
  pivot_longer(cols=cnames[intervals], names_to="colname", values_to="value") %>%
  ggplot() +
  geom_boxplot(aes(y=value)) +
  facet_wrap(vars(colname), scales="free")
```



```r
# Outliers of interval data are removed
df_main %<>%
  remove_outlier(columns=cnames[intervals])
# Plotting outliers for interval data after removal
df_main %>%
  pivot_longer(cols=cnames[intervals], names_to="colname", values_to="value") %>%
  ggplot() +
  geom_boxplot(aes(y=value)) +
  facet_wrap(vars(colname), scales="free")
```




```r
# As an initial evalutation of association between data, the correlation matrix of interval data are calculated

initcor <- cor(x=df_main[intervals], use="complete.obs", method="pearson")
# A heatmap of correlation matrix
corrplot(initcor, method = 'square', diag = FALSE, order = 'hclust',
         addrect = 3, rect.col = 'blue', rect.lwd = 3, tl.pos = 'd', tl.cex = 0.7)
```



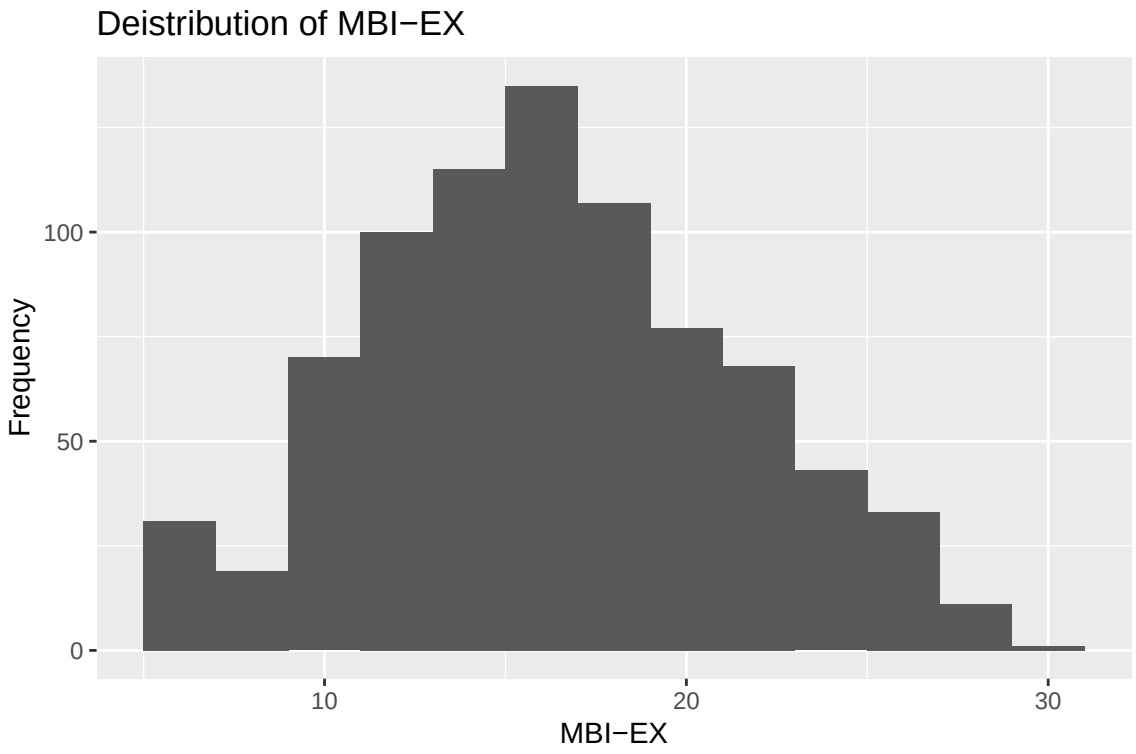

```
df_focus <- df_main %>% select(mbi_ex, cesd)

stat_mbi_ex <- stat.desc(df_focus$mbi_ex, norm=TRUE)
stat_cesd <- stat.desc(df_focus$cesd, norm=TRUE)
stat_mbi_ex
```

```
##       nbr.val      nbr.null        nbr.na           min           max
##  8.100000e+02  0.000000e+00  0.000000e+00  5.000000e+00  3.000000e+01
##         range           sum        median          mean       SE.mean
##  2.500000e+01  1.359700e+04  1.700000e+01  1.678642e+01  1.784702e-01
##  CI.mean.0.95           var       std.dev      coef.var      skewness
##  3.503193e-01  2.579982e+01  5.079352e+00  3.025870e-01  1.128473e-01
##      skew.2SE      kurtosis      kurt.2SE    normtest.W    normtest.p
##  6.567956e-01 -3.932815e-01 -1.145891e+00  9.897747e-01  2.021627e-05
```

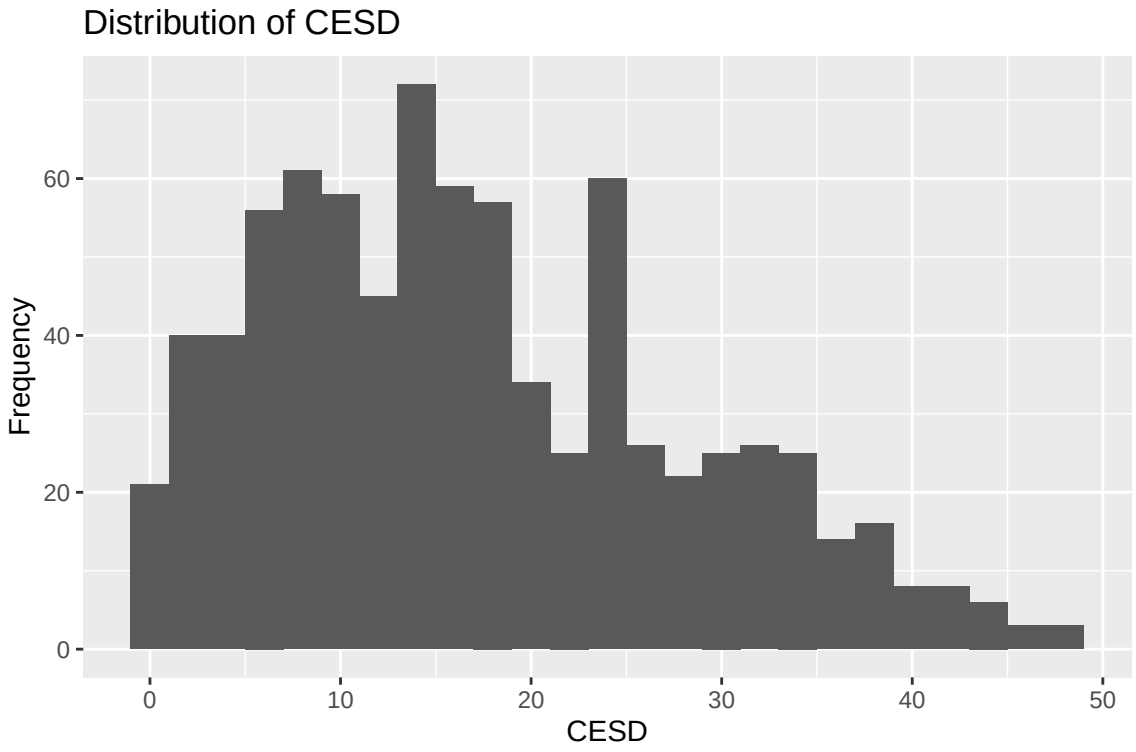
```
stat_cesd
```

```
##       nbr.val      nbr.null        nbr.na           min           max
##  8.100000e+02  9.000000e+00  0.000000e+00  0.000000e+00  4.900000e+01
##         range           sum        median          mean       SE.mean
##  4.900000e+01  1.425400e+04  1.600000e+01  1.759753e+01  3.818207e-01
##  CI.mean.0.95           var       std.dev      coef.var      skewness
##  7.494761e-01  1.180875e+02  1.086681e+01  6.175189e-01  5.622770e-01
##      skew.2SE      kurtosis      kurt.2SE    normtest.W    normtest.p
##  3.272573e+00 -3.914713e-01 -1.140617e+00  9.617336e-01  1.031039e-13
```

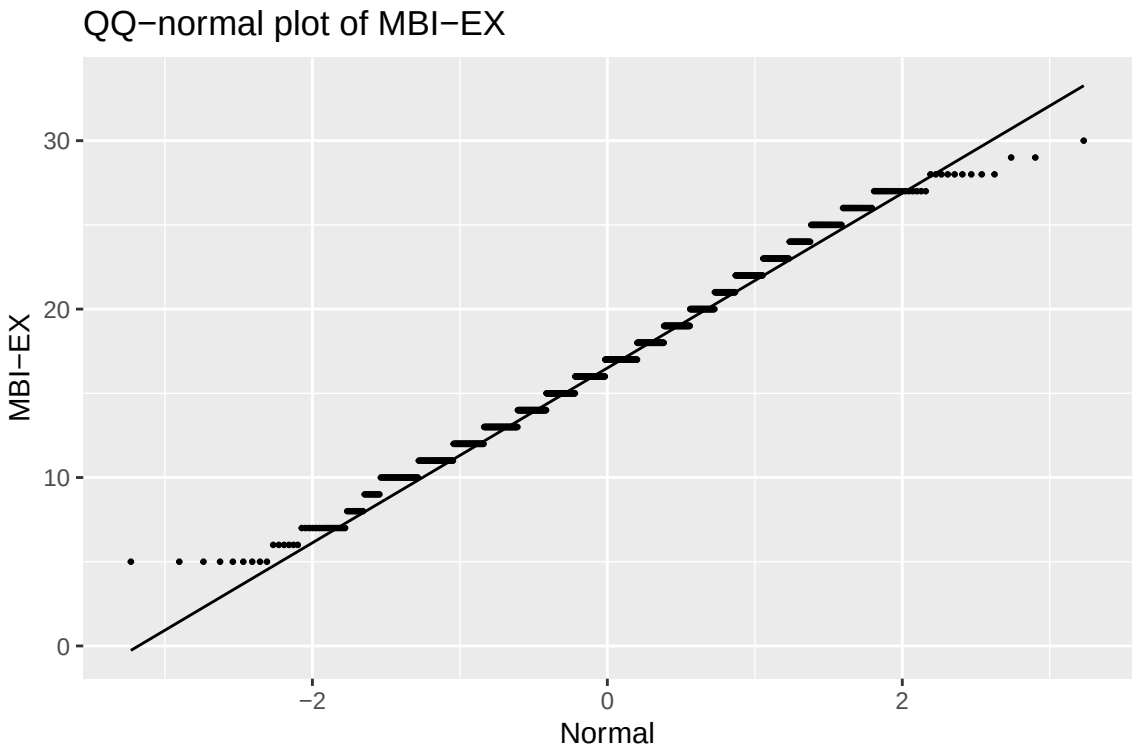
```
# First histogram and qq normal plot are generated to explore the overal
# distribution and normality of the data.
ggplot(df_focus) +
  geom_histogram(aes(x=mbi_ex),binwidth=2) +
  labs(x="MBI-EX", y="Frequency", title="Deistribution of MBI-EX")
```



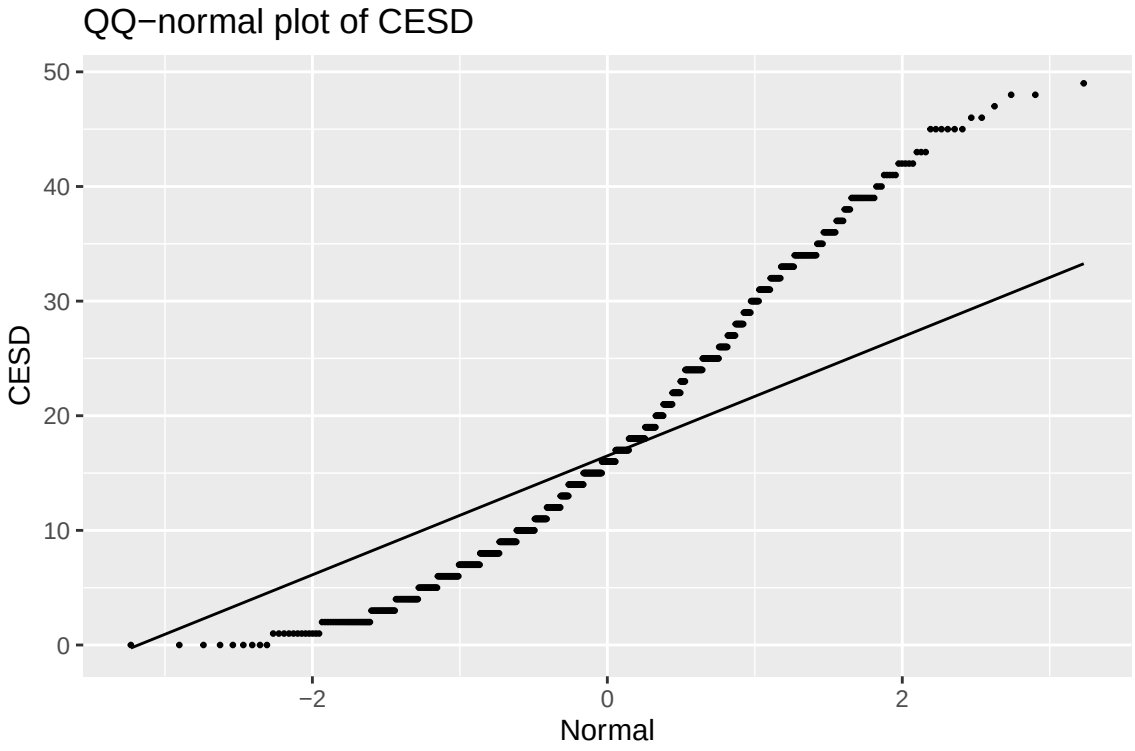

```
ggplot(df_focus) +
  geom_histogram(aes(x=cesd),binwidth=2) +
  labs(x="CESD", y="Frequency", title="Distribution of CESD")
```




```
ggplot(df_focus) +
  geom_qq(aes(sample=mbi_ex), size=0.5) +
  stat_qq_line(aes(sample=mbi_ex)) +
  labs(x="Normal", y="MBI-EX", title="QQ-normal plot of MBI-EX")
```

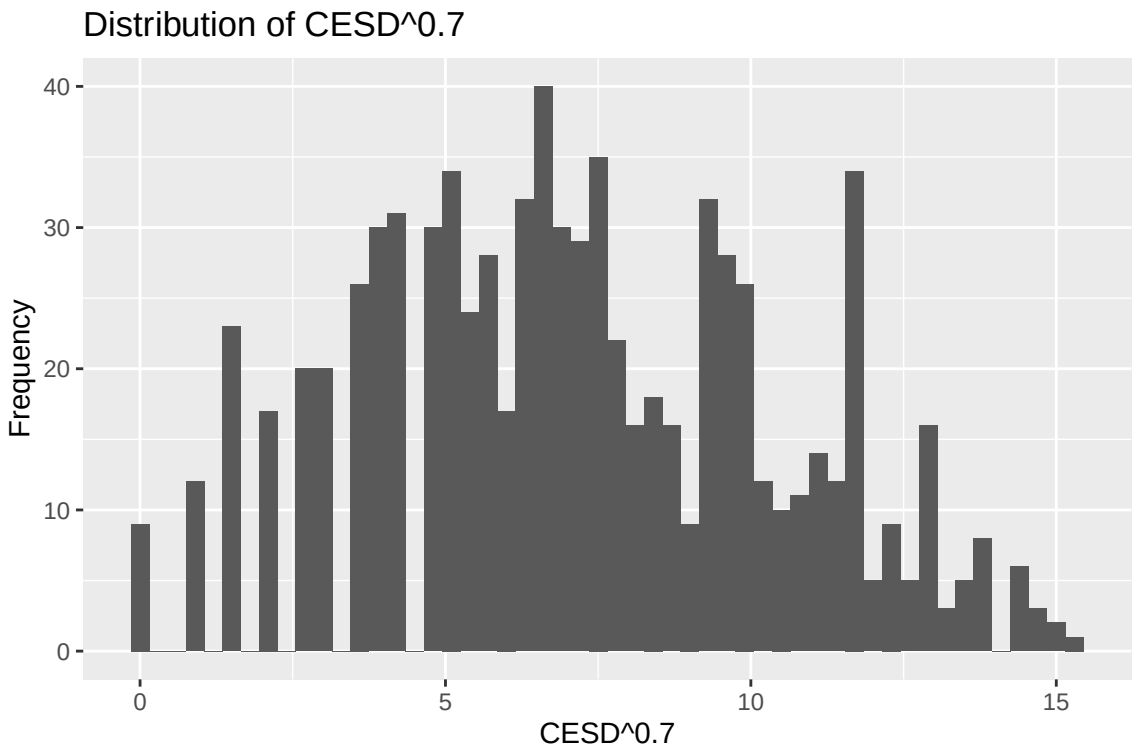



```
ggplot(df_focus) +
  geom_qq(aes(sample=cesd), size=0.5) +
  stat_qq_line(aes(sample=mbi_ex)) +
  labs(x="Normal", y="CESD", title="QQ-normal plot of CESD")
```

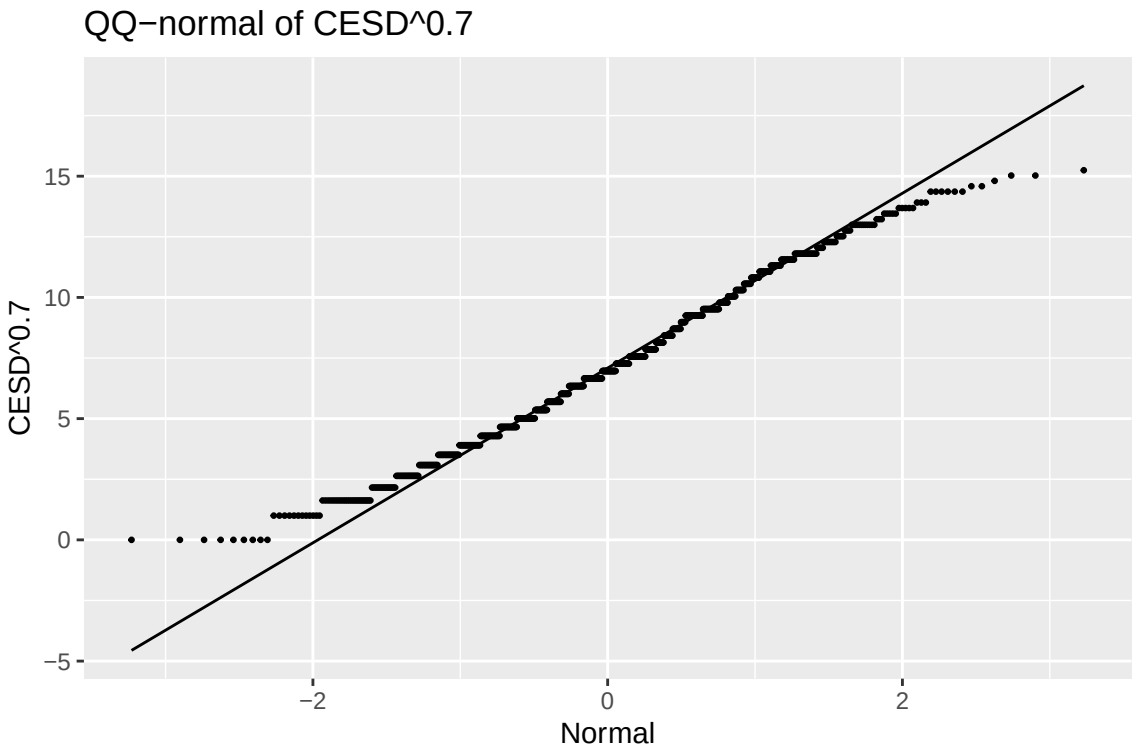



```
# Tranformation function
tr <- function(x)x^0.7
tr_inv <- function(y)y^(10/7)
```

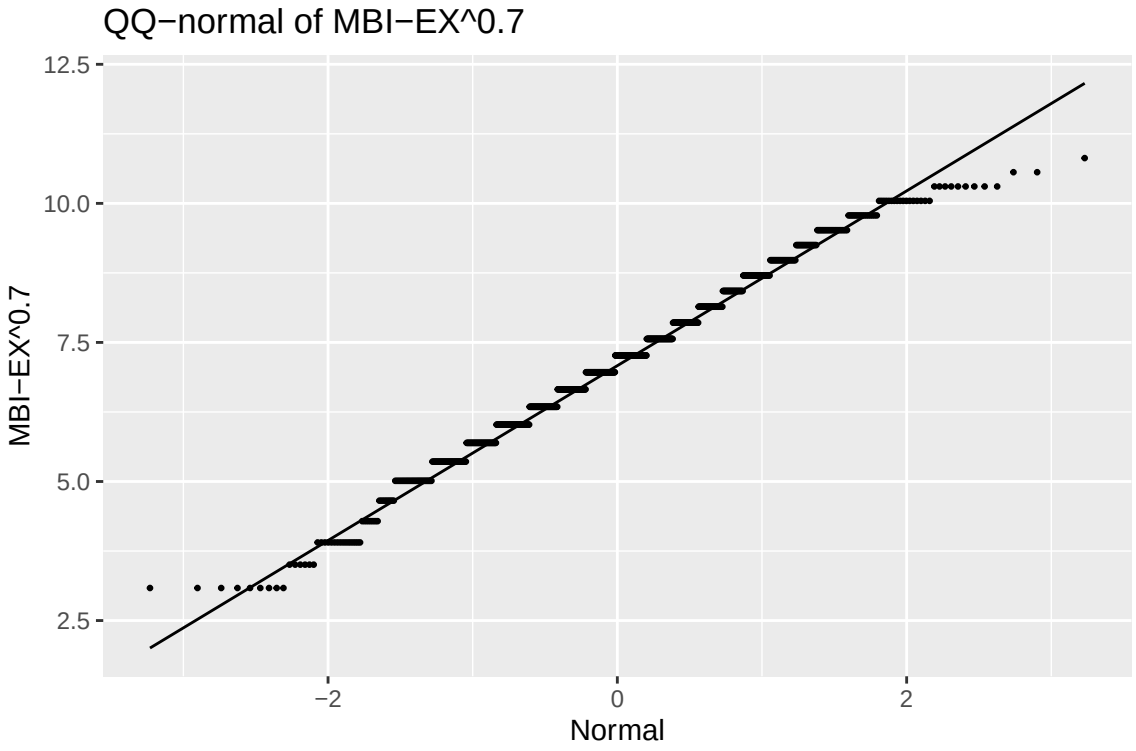
```r
# Transforming data to improve normality
df_focus_tr <- df_focus %>%
  mutate(cesd_tr=tr(cesd), mbi_ex_tr=tr(mbi_ex))

df_focus_tr %>%
  ggplot() +
  geom_histogram(aes(x=cesd_tr), binwidth=0.3)+
  labs(x="CESD^0.7", y="Frequency", title="Distribution of CESD^0.7")
```




```r
ggplot(df_focus_tr) +
  geom_qq(aes(sample=cesd_tr), size=0.5) +
  stat_qq_line(aes(sample=cesd_tr)) +
  labs(x="Normal", y="CESD^0.7", title="QQ-normal of CESD^0.7")
```




```
ggplot(df_focus_tr) +
  geom_qq(aes(sample=mbi_ex_tr), size=0.5) +
  stat_qq_line(aes(sample=mbi_ex_tr)) +
  labs(x="Normal", y="MBI-EX^0.7", title="QQ-normal of MBI-EX^0.7")
```




```
stat_mbi_ex_tr <- stat.desc(df_focus_tr$mbi_ex_tr, norm=TRUE)
stat_cesd_tr <- stat.desc(df_focus_tr$cesd_tr, norm=TRUE)
stat_mbi_ex_tr
```

```
##       nbr.val      nbr.null        nbr.na           min           max
##  8.100000e+02  0.000000e+00  0.000000e+00  3.085169e+00  1.081396e+01
##         range           sum        median          mean       SE.mean
##  7.728794e+00  5.774791e+03  7.266315e+00  7.129371e+00  5.437560e-02
##  CI.mean.0.95           var       std.dev      coef.var      skewness
##  1.067339e-01  2.394932e+00  1.547557e+00  2.170678e-01 -1.337591e-01
##      skew.2SE      kurtosis      kurt.2SE    normtest.W    normtest.p
## -7.785069e-01 -2.608577e-01 -7.600525e-01  9.901356e-01  2.949912e-05
```

```
stat_cesd_tr
```

```
##       nbr.val      nbr.null        nbr.na           min           max
##  8.100000e+02  9.000000e+00  0.000000e+00  0.000000e+00  1.524534e+01
##         range           sum        median          mean       SE.mean
##  1.524534e+01  5.760359e+03  6.964405e+00  7.111554e+00  1.154262e-01
##  CI.mean.0.95           var       std.dev      coef.var      skewness
##  2.265702e-01  1.079180e+01  3.285088e+00  4.619368e-01  1.486183e-01
##      skew.2SE      kurtosis      kurt.2SE    normtest.W    normtest.p
##  8.649904e-01 -6.134731e-01 -1.787456e+00  9.885245e-01  5.715625e-06
```

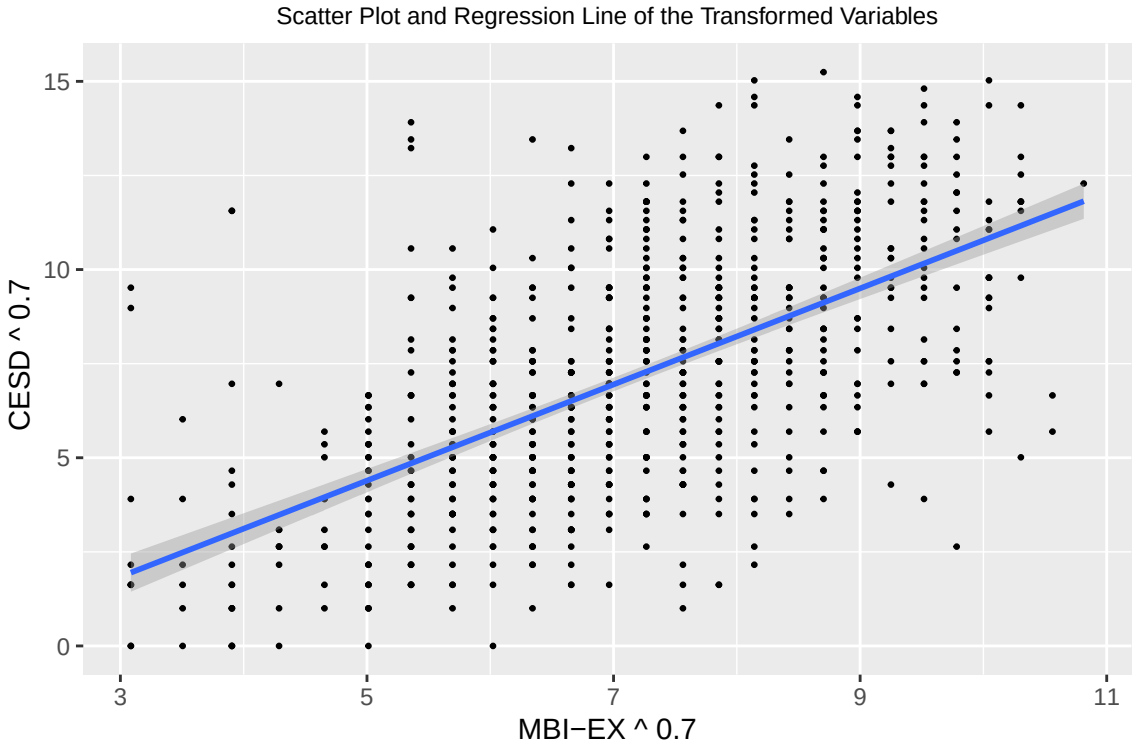
```
cortest_pearson <- cor.test(df_focus_tr$mbi_ex_tr, df_focus_tr$cesd_tr, method="pearson")
cortest_spearman <- cor.test(df_focus_tr$mbi_ex, df_focus_tr$cesd, method="spearman")
```

```
## Warning in cor.test.default(df_focus_tr$mbi_ex, df_focus_tr$cesd, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
bootTau <- function(df, i) cor(df$mbi_ex[i], df$cesd[i],
                               method="spearman", use="complete.obs")
boot_spearman <- boot(df_focus, bootTau, 1000)
spearman_boot_ci <- boot.ci(boot_spearman)
```

```
## Warning in boot.ci(boot_spearman): bootstrap variances needed for studentized
## intervals
```

```
df_focus_tr_long <- df_focus_tr %>%
  pivot_longer(cols=c(mbi_ex_tr,cesd_tr), names_to="vars", values_to="vals")
levtest <- leveneTest(df_focus_tr_long$vals, as.factor(df_focus_tr_long$vars))

df_focus_tr %>%
  ggplot(aes(x=mbi_ex_tr, y=cesd_tr)) +
  geom_point(size=0.5) +
  geom_smooth(formula = y ~ x, method=lm) +
  labs(x="MBI-EX ^ 0.7", y=" CESD ^ 0.7",
       title="Scatter Plot and Regression Line of the Transformed Variables") +
  theme(plot.title = element_text(size = 9, hjust=0.5))
```


Scatter Plot and Regression Line of the Transformed Variables
CESD ^ 0.7
MBI−EX ^ 0.7
0
5
10
15
3
5
7
9
11