Analysis of PISA data using corVis

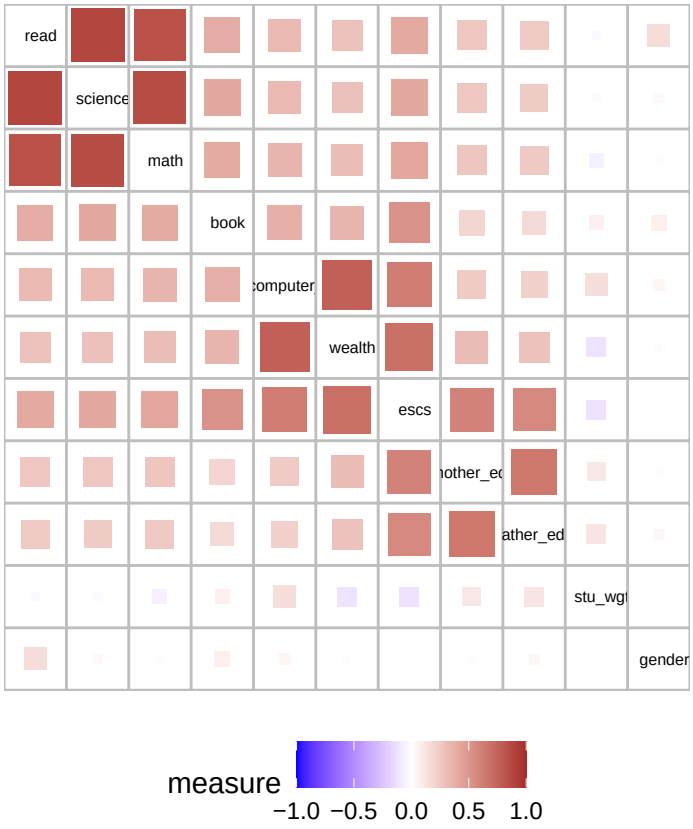
Amit Chinwan

The R package *learningtower* provides Programme for International Student Assessment (PISA) data from OECD. PISA is an international assessment measuring student performance in reading, mathematical and scientific literacy. The data is collected for 15 year olds on a three year basis and is available for the years 2000 - 2018. We use 2018 student data to show how corVis can be implemented to detect patterns.

```
data(student_subset_2018)
df <- student_subset_2018

# removing variables such as school_id, student_id
df$school_id <- NULL
df$student_id <- NULL

# removing country for this analysis
df$country <- NULL
df$year <- NULL

#removing variables related to wealth
df$dishwasher <- NULL
df$desk <- NULL
df$car <- NULL
df$room <- NULL
df$computer <- NULL
df$internet <- NULL
df$television <- NULL

# making mother education, father education, number of computers, number of books ordinal
df$mother_educ <- ordered(df$mother_educ, levels=c("less than ISCED1",
                                                   "ISCED 1",
                                                   "ISCED 2",
                                                   "ISCED 3A",
                                                   "ISCED 3B, C" ))
df$father_educ <- ordered(df$father_educ, levels=c("less than ISCED1",
                                                   "ISCED 1",
                                                   "ISCED 2",
                                                   "ISCED 3A",
                                                   "ISCED 3B, C" ))
df$computer_n <- as.ordered(df$computer_n)
df$book <- ordered(df$book, levels=c("0-10", "11-25", "26-100", "101-200", "201-500",
                                     "more than 500"))
```

Association matrix display

```
a <- calc_assoc(df)
plot_assoc_matrix(a)
```



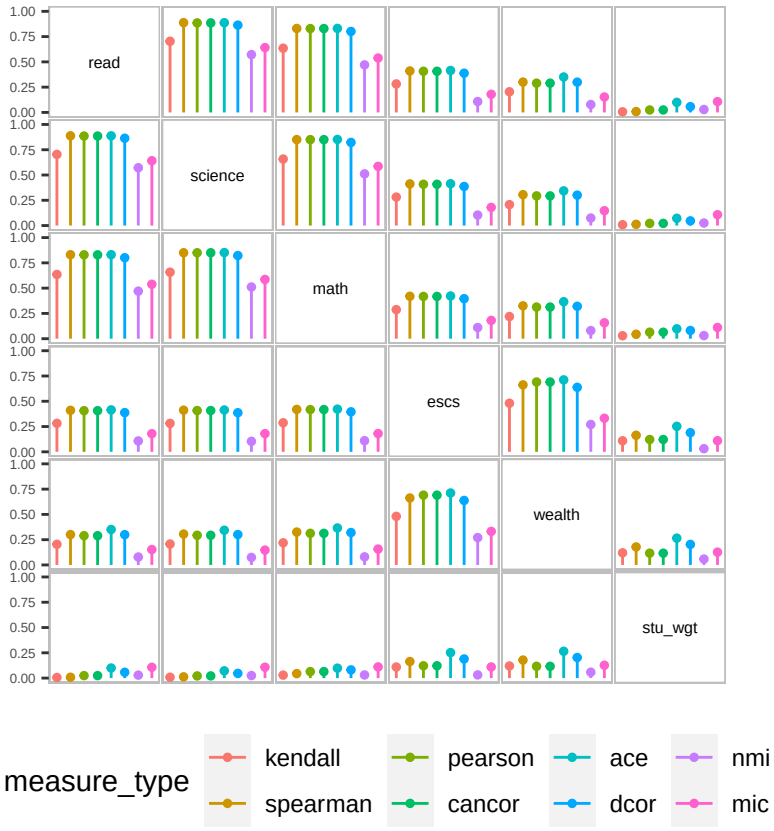
The plot shows a strong positive correlation between reading, science and maths score for students. These scores are highly associated with the number of books a student has. A high association between wealth, escs and computer_n suggests that higher wealth lead to a better social and cultural status.

The negative correaltion between wealth and stu_wgt suggests that there are students who does well overall with poor socio-economic status.

Multiple measures display

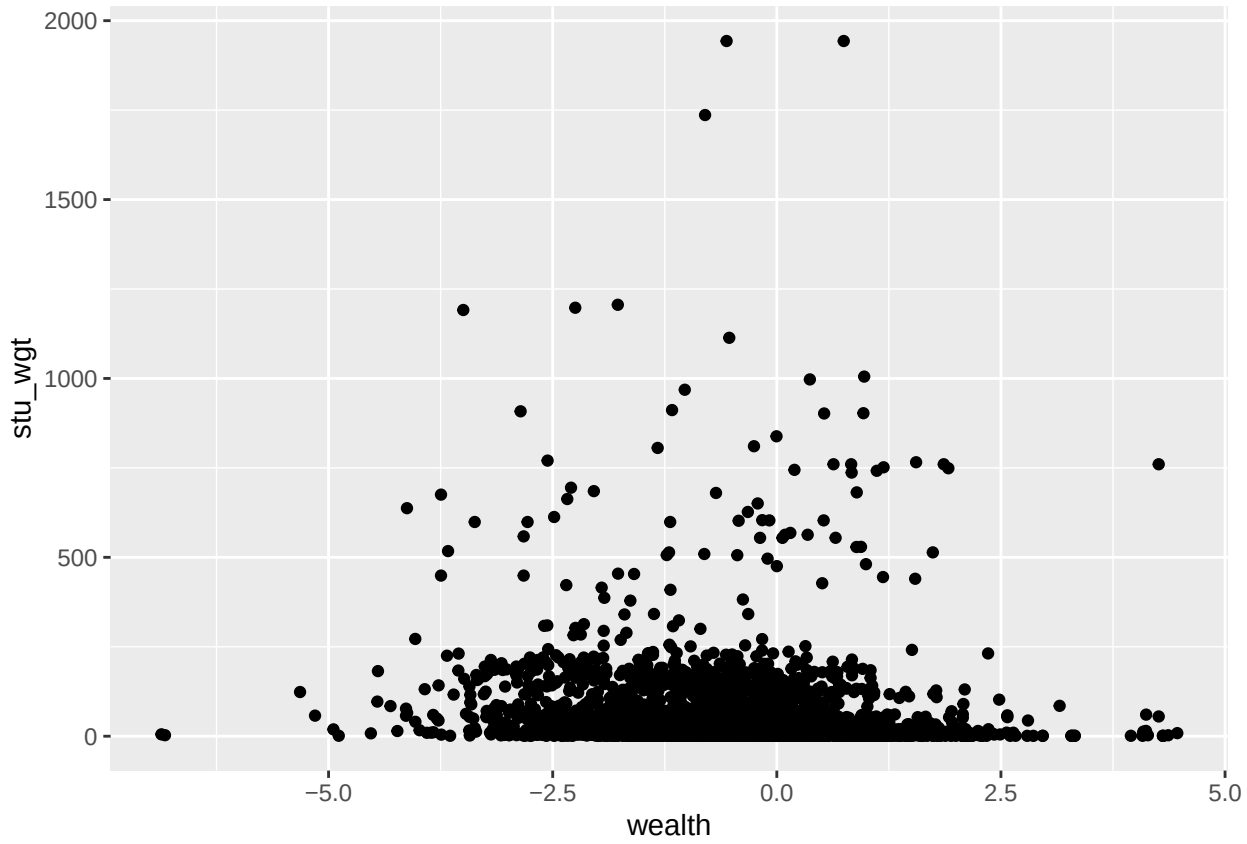
We select only numeric variables from the dataset and use multiple measures for a comparison.

```
df_num <- dplyr::select(df,where(is.numeric))
c <- calc_assoc_all(df_num)
plot_assoc_matrix(c)
```

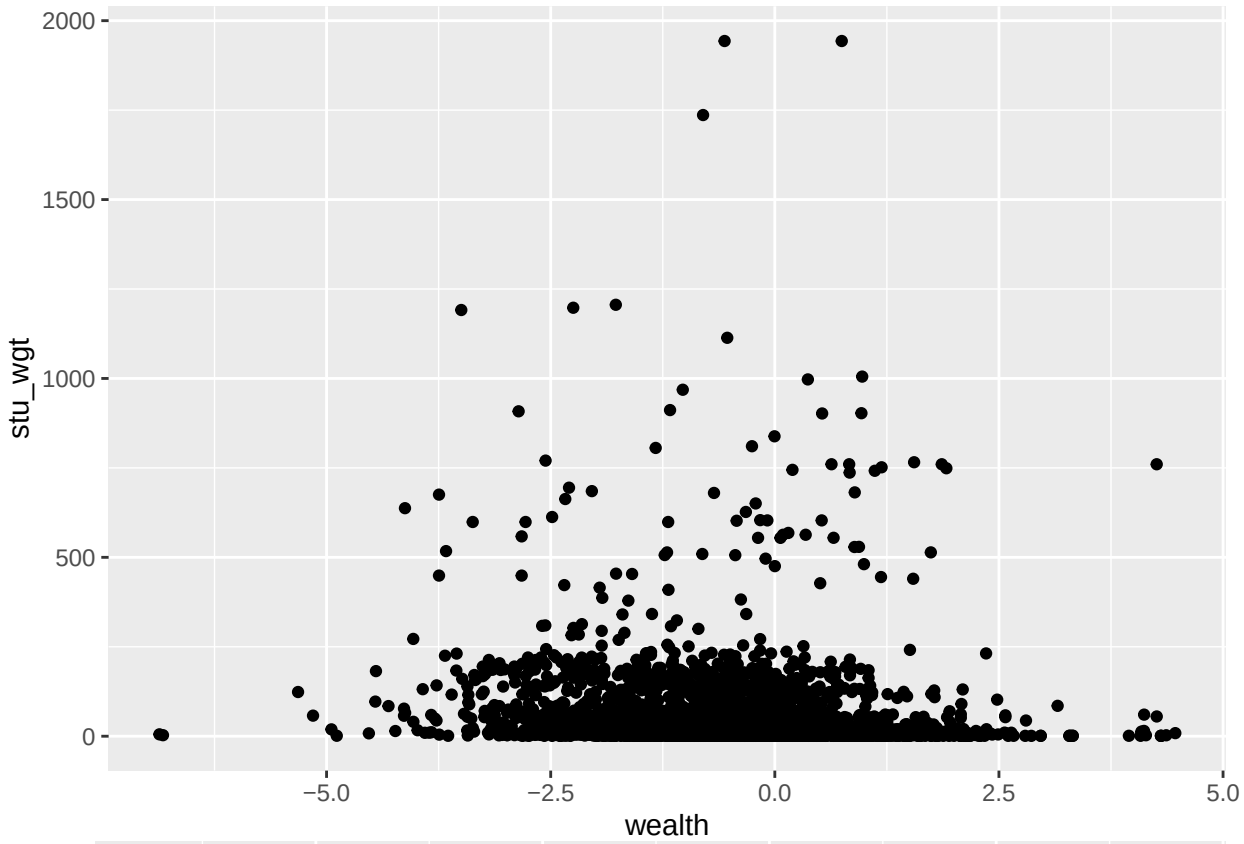


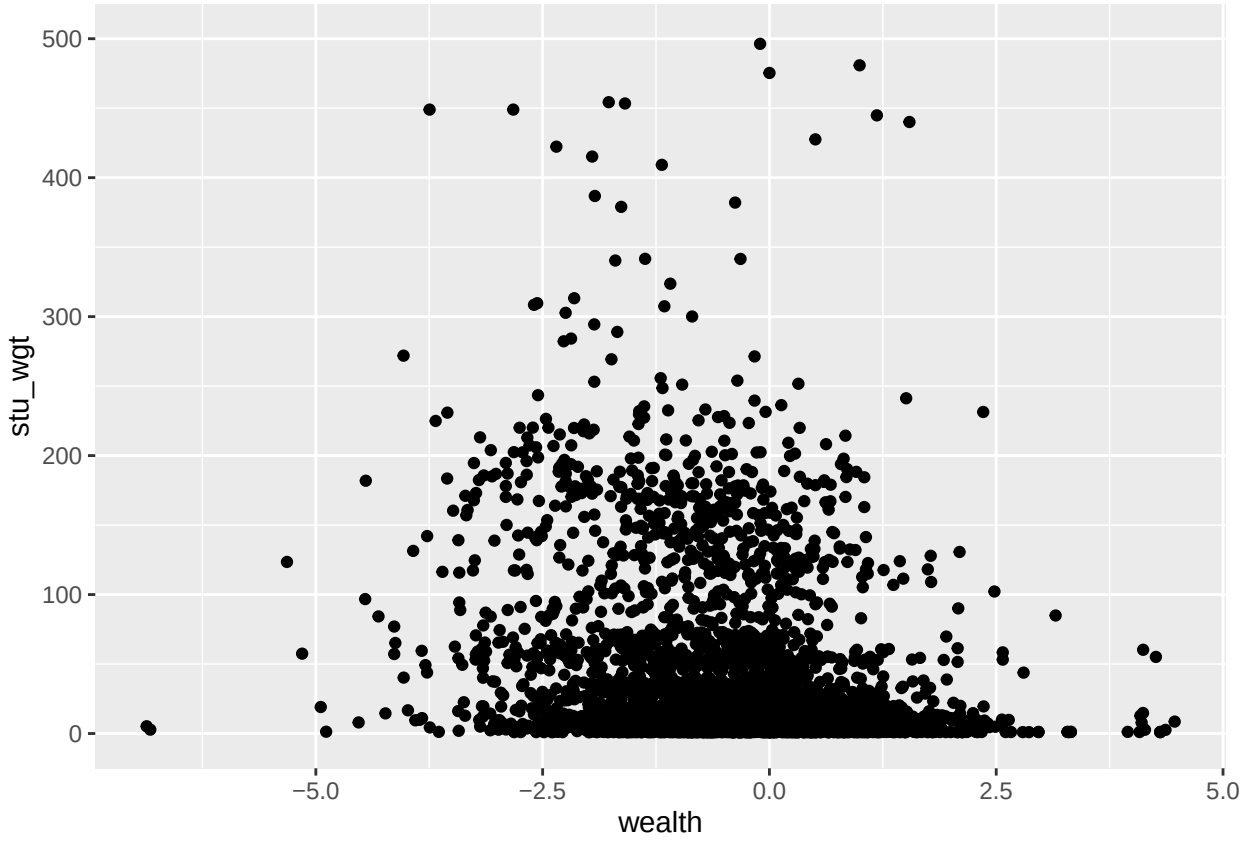
The variable pair (wealth, stu_wgt) is interesting as all the measures tend to be low compared to ace and dcor. If we look at more closely, there seems to be a non-linear relationship between these variables which ace and dcor capture quite efficiently.

```
show_assoc(df_num,"wealth","stu_wgt")
```



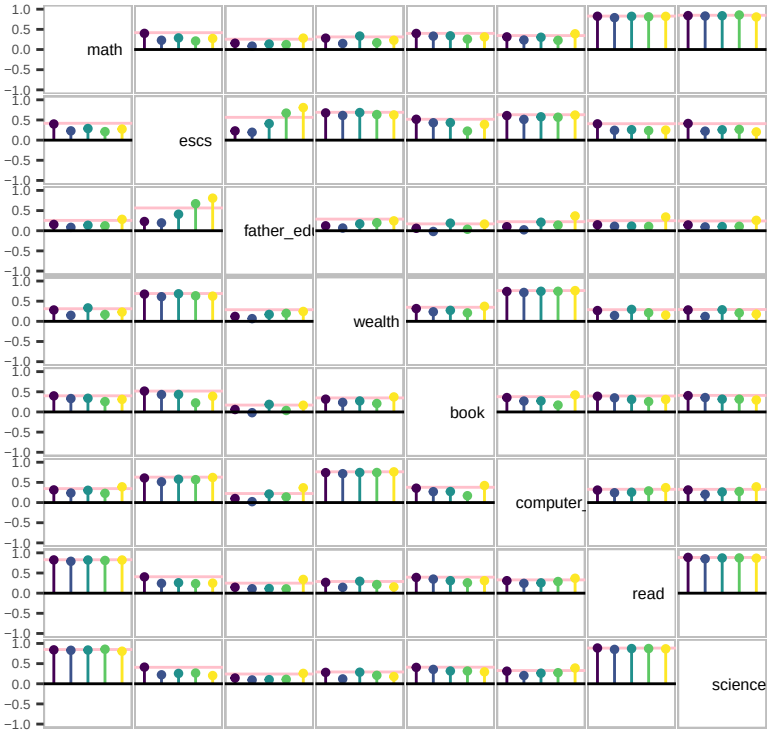

```
show_assoc(df_num,"wealth","stu_wgt") + ggplot2::ylim(0,500)
```


2000
1500
1000
500
0
stu_wgt
−5.0
−2.5
0.0
2.5
5.0
wealth


500
400
300
200
100
0
stu_wgt
−5.0
−2.5
0.0
2.5
5.0
wealth

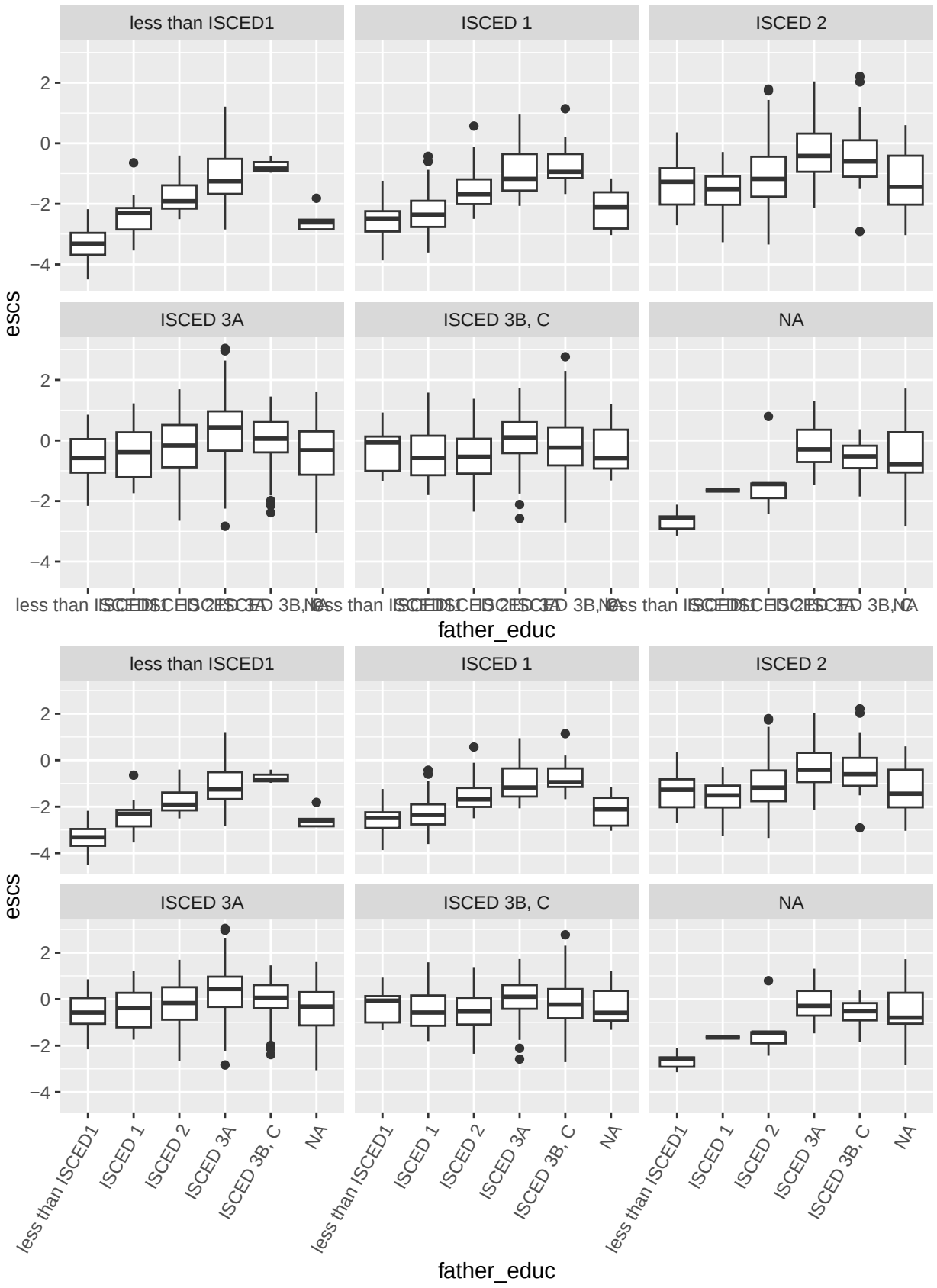
Conditional Assocaition plot

```
# removing gender and stu_wgt as having weak association when father_edu is used as by variable
df_new <- df
df_new$gender <- NULL
df_new$stu_wgt <- NULL

b_new <- calc_assoc(df_new, by="mother_educ")
plot_assoc_matrix(b_new)
```



The plot shows a strong association for variable pair (escs,father_educ) for different levels of mother's education. This can be seen more clearly from the boxplot below.

```
show_assoc(df_new,"father_educ","escs","mother_educ") +
  ggplot2::theme(axis.text.x = ggplot2::element_text(angle = 60, hjust=1))
```


less than ISCED1
ISCED 1
ISCED 2
ISCED 3A
ISCED 3B, C
NA
escs
2
0
−2
−4
less than ISCED1
ISCED 1
ISCED 2
ISCED 3A
ISCED 3B, C
NA
father_educ