COS-D419 Factor Analysis and Structural Equation Models 2023, Assignment 2

Rong Guang

2023-01-30

1 Exercise 2.1

Specify and test the hypothesis given on the page 1 of the lecture material.

Draw conclusions based on the χ^2 statistic and the CFI, TLI, RMSEA, and SRMR indices.

What can you say about the parameter estimates?

Visualize the model.

1.1 Read in the data set

Start by downloading the data file from Moodle to Project folder.

```r
library(tidyverse)#data wrangling
library(readr)# read data into r
orig_data <- read_csv("ASC7INDM.CSV", show_col_types = FALSE)
```

1.2 Write functions

```r
unique.levels <-  function(sc){
  values <- lapply(sc, function(x)sort(unique(x)))
for(x in 1:ncol(sc)){
  a <- paste(c("Variable ",
               names(values)[x],
               " has values of ",
               paste(values[[x]],
                     collapse = ",")),
             collapse = "")
  print(a)
  }
}
```

1.3 Subset the data set

Subset the variables for analysis and name it as sc (Self-concept).

```
# Select the variables for use
sc <- orig_data %>% dplyr::select(starts_with("SDQ2N")) # namming logic: sc = self-concept
```

1.4 Inspect the data

Have a quick overview of the data.

```
glimpse(sc)
```

```
## Rows: 265
## Columns: 16
## $ SDQ2N01 <dbl> 6, 6, 4, 5, 6, 5, 1, 2, 5, 4, 2, 5, 6, 4, 4, 6, 6, 6, 5, 6, 6,~
## $ SDQ2N13 <dbl> 5, 6, 6, 5, 5, 5, 6, 1, 5, 6, 6, 5, 6, 3, 5, 6, 6, 6, 4, 5, 5,~
## $ SDQ2N25 <dbl> 4, 6, 6, 5, 5, 5, 1, 6, 6, 3, 6, 6, 6, 5, 5, 6, 6, 6, 6, 5, 4,~
## $ SDQ2N37 <dbl> 6, 6, 2, 6, 4, 3, 6, 4, 6, 6, 6, 5, 5, 5, 4, 5, 6, 4, 4, 6, 6,~
## $ SDQ2N04 <dbl> 3, 6, 6, 5, 3, 3, 4, 4, 6, 6, 5, 6, 5, 4, 4, 4, 4, 6, 5, 5, 3,~
## $ SDQ2N16 <dbl> 4, 6, 4, 6, 4, 2, 6, 4, 6, 5, 6, 6, 5, 5, 5, 5, 6, 5, 4, 6, 6,~
## $ SDQ2N28 <dbl> 4, 6, 6, 5, 4, 4, 6, 4, 6, 6, 6, 6, 5, 5, 5, 5, 6, 4, 2, 4, 4,~
## $ SDQ2N40 <dbl> 6, 6, 3, 6, 4, 4, 6, 6, 6, 6, 6, 6, 6, 5, 4, 4, 6, 6, 5, 5, 5,~
## $ SDQ2N10 <dbl> 2, 5, 6, 5, 4, 4, 1, 6, 5, 4, 2, 6, 5, 5, 5, 3, 4, 6, 5, 4, 6,~
## $ SDQ2N22 <dbl> 6, 6, 5, 6, 6, 4, 6, 6, 6, 6, 6, 6, 6, 5, 6, 6, 6, 6, 6, 3, 6,~
## $ SDQ2N34 <dbl> 1, 6, 4, 3, 5, 5, 1, 1, 5, 4, 5, 6, 5, 2, 5, 2, 3, 2, 1, 3, 3,~
## $ SDQ2N46 <dbl> 5, 6, 5, 5, 6, 6, 6, 5, 6, 6, 6, 6, 6, 6, 2, 5, 6, 6, 6, 6, 6,~
## $ SDQ2N07 <dbl> 6, 6, 6, 6, 3, 4, 5, 3, 6, 5, 6, 6, 6, 6, 4, 4, 6, 6, 6, 6, 3,~
## $ SDQ2N19 <dbl> 6, 6, 6, 6, 4, 5, 6, 4, 6, 6, 5, 6, 6, 6, 5, 5, 6, 6, 5, 5, 5,~
## $ SDQ2N31 <dbl> 6, 6, 3, 6, 4, 4, 6, 4, 6, 6, 6, 6, 6, 6, 5, 5, 6, 6, 5, 5, 5,~
## $ SDQ2N43 <dbl> 6, 6, 1, 5, 5, 4, 5, 6, 6, 6, 6, 6, 6, 6, 5, 6, 6, 6, 5, 6, 5,~
```

The data set includes 16 variables from 265 observations. All the variables are numeric. Next, I examined the unique values of each variables.

```
unique.levels(sc)
```

```
## [1] "Variable SDQ2N01 has values of 1,2,3,4,5,6"
## [1] "Variable SDQ2N13 has values of 1,2,3,4,5,6"
## [1] "Variable SDQ2N25 has values of 1,2,3,4,5,6"
## [1] "Variable SDQ2N37 has values of 1,2,3,4,5,6"
## [1] "Variable SDQ2N04 has values of 1,2,3,4,5,6"
## [1] "Variable SDQ2N16 has values of 1,2,3,4,5,6"
## [1] "Variable SDQ2N28 has values of 1,2,3,4,5,6"
## [1] "Variable SDQ2N40 has values of 1,2,3,4,5,6"
## [1] "Variable SDQ2N10 has values of 1,2,3,4,5,6"
## [1] "Variable SDQ2N22 has values of 1,2,3,4,5,6"
## [1] "Variable SDQ2N34 has values of 1,2,3,4,5,6"
## [1] "Variable SDQ2N46 has values of 1,2,3,4,5,6"
## [1] "Variable SDQ2N07 has values of 1,2,3,4,5,6"
## [1] "Variable SDQ2N19 has values of 1,2,3,4,5,6"
## [1] "Variable SDQ2N31 has values of 1,2,3,4,5,6"
## [1] "Variable SDQ2N43 has values of 1,2,3,4,5,6"
```

For each variable, the values distribute from 1 to 6.

2 Explore the data

2.1 Descriptive statistics

```
library(kableExtra)#improved table visuals
library(psych)#for function "describe"
sc.ds <- sc %>%  #sc.ds = self-concept descriptive statistics
  describe(IQR = T) %>%
    as.data.frame() %>%
  dplyr::select(mean, median, sd, range, se, IQR)
#print the descriptive statistics table
sc.ds %>%
  kable(booktabs=T,
        longtable=T,
        digits = 2,
        caption = "Descriptive dtatistics of selected variables",
        linesep = "") %>%
  add_header_above(c("", "centralized tendency" = 2, "dispersion tendency" = 4)) %>%
  kable_styling(latex_options = c("striped","repeat_header")) %>%
  column_spec(1, width = "3cm", bold = T, color = "red")
```

Table 1: Descriptive dtatistics of selected variables

	centralized tendency		dispersion tendency			
	mean	median	sd	range	se	IQR
SDQ2N01	4.41	5	1.35	5	0.08	1
SDQ2N13	5.00	6	1.36	5	0.08	2
SDQ2N25	5.10	6	1.23	5	0.08	1
SDQ2N37	4.83	5	1.14	5	0.07	2
SDQ2N04	4.52	5	1.40	5	0.09	2
SDQ2N16	4.65	5	1.24	5	0.08	2
SDQ2N28	4.69	5	1.33	5	0.08	2
SDQ2N40	4.98	5	1.36	5	0.08	1
SDQ2N10	4.62	5	1.15	5	0.07	1
SDQ2N22	5.38	6	1.09	5	0.07	1
SDQ2N34	3.89	4	1.70	5	0.10	3
SDQ2N46	5.27	6	1.30	5	0.08	1
SDQ2N07	4.32	5	1.78	5	0.11	3
SDQ2N19	4.54	5	1.69	5	0.10	2
SDQ2N31	4.74	5	1.57	5	0.10	2
SDQ2N43	4.98	5	1.40	5	0.09	1

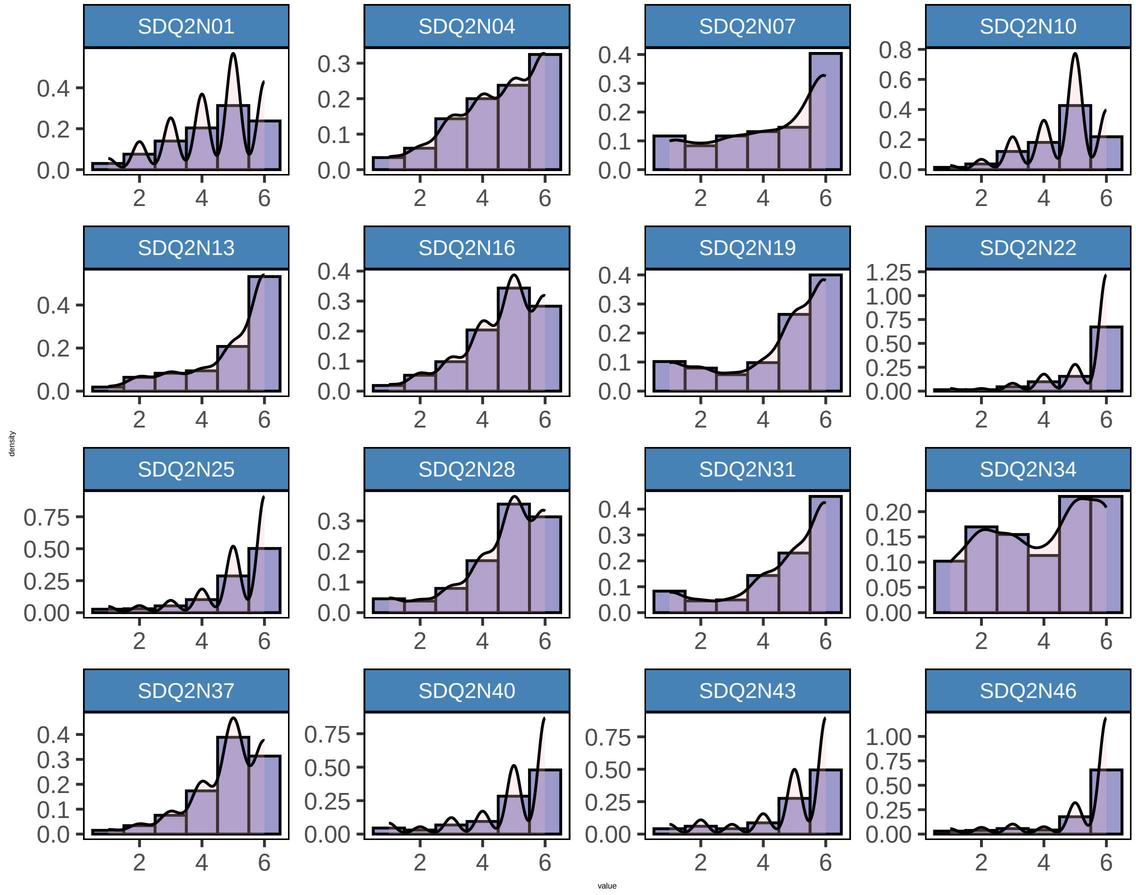
```
#sc.ds %>%
#  tab_df(digits = 2,
#        alternate.rows = T,
#        title = "Table 1. Descriptive dtatistics of selected variables",
#        CSS = list(css.centralign='text-align: right;'))
```

2.2 Visualization

2.2.1 Histogram

```
sc %>%
  pivot_longer(everything()) %>%  #longer format
  ggplot(aes(x = value)) + #x axis used variable "value" (a default of pivot)
  geom_histogram(binwidth = 1, aes(y = ..density..), #match ys of density and histogram plots
                 color = "black",  fill = "#9999CC")+  # adjust aesthetics for hist
  geom_density(fill = "pink", alpha = 0.25)+ #adjust aesthetics for density plot
  facet_wrap(~name, scales = "free") + #wrap by name variable
  theme(panel.grid.major = element_blank(), #get rid of the  grids
        panel.grid.minor = element_blank(),
        panel.background = element_rect(fill = "white",#adjust the background
                                        color = "black"),
        strip.background = element_rect(color = "black",#adjust the strips aes
                                        fill = "steelblue"),
        strip.text = element_text(size =8, color = "white"), #adjust strip text
        axis.title.x = element_text(size = 3), #adjust the x text
        axis.title.y = element_text(size = 3), # adjust the y text
        plot.title = element_text(size = 12, face = "bold"))+ #adjust the title
  labs(title = "Figure 1 Distribution of selected items") #title it
```

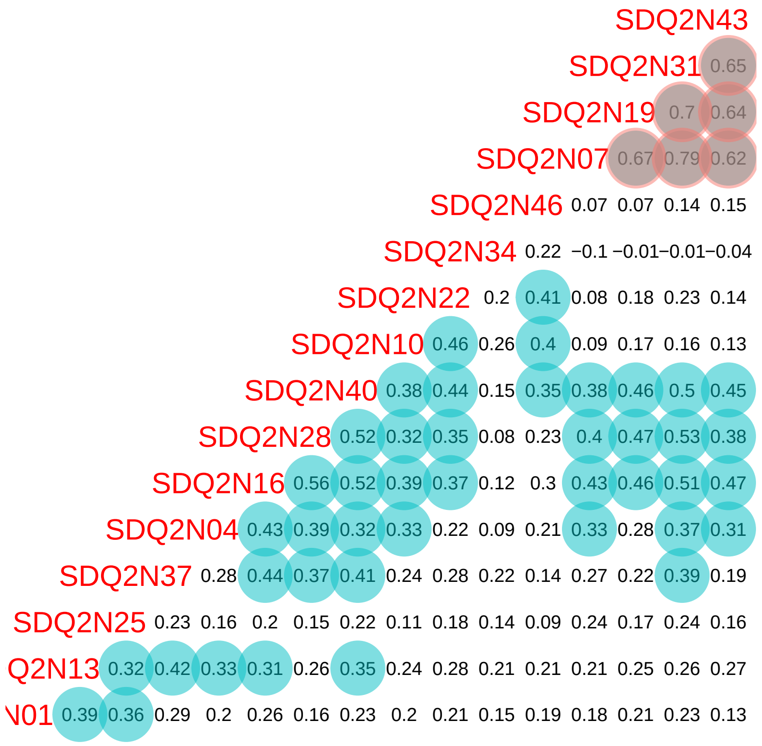
Figure 1 Distribution of selected items



2.2.2 Correlation plot

```
library(GGally)
ggcorr(sc,
       geom = "blank",
       label = TRUE,
       hjust = 0.85,
       color = "red",
       face = "bold",
       method = c("pairwise","spearman"),
       digits = 2,
       label_size = 2.5,
       label_round = 2) +
  geom_point(size = 9,
             aes(color = "red",
                 alpha = abs(coefficient) > 0.3)) +
  scale_alpha_manual(values = c("TRUE" = 0.3, "FALSE" = 0)) +
    geom_point(size = 10,
               aes(color = "green", alpha = abs(coefficient) > 0.6)) +
  scale_alpha_manual(values = c("TRUE" = 0.5, "FALSE" = 0)) +
  guides(color = FALSE,
         alpha = FALSE) +
  labs(title = "Figure 2. Spearman correlation matrix of the selected items",
       caption =
         "Red circles indicates correlation coefficient >= 0.5; gree circle indicates >= 0.3")
```

Figure 2. Spearman correlation matrix of the selected items



Red circles indicates correlation coefficient >= 0.5; gree circle indicates >= 0.3

It is found that each variable correlated with at least one of the other variable with a spearman correlation coefficient >= 0.3, except for item SDQ2N46 and SDQ2N34.