MA615project

Yichu Yan

10/18/2019

WVS WAVE 7

In 2015 WVSA started planning the 7th wave to be conducted worldwide in 2017-2020. Subsequent waves are planned every five years.

Strategic goals for the 7th wave include: - Expansion of territorial coverage from 60 countries in WVS 6 to 80 in WVS 7; - Deepening collaboration within the international development community; - Deepening collaboration within NGOs, academic institutions & research foundations; - Updating the WVS-7 questionnaire with new topics & items covering new social phenomena and emerging processes of value change; - Expanding the 7th wave WVS with data useful for monitoring the SDGs; - Expanding capacity and resources for survey fieldwork in developing countries.

The 7th wave will continue monitoring cultural values, attitudes and beliefs towards gender, family, and religion; attitudes and experience of poverty; education, health, and security; social tolerance and trust; attitudes towards multilateral institutions; cultural differences and similarities between regions and societies. In addition, the WVS-7 questionnaire has been elaborated with the inclusion of such new topics as the issues of justice, moral principles, corruption, accountability and risk, migration, national security and global governance.

Load Data

Select following survey questions

q1 - V10: Feeling of happiness
q2 - V11: State of health (subjective)
q3 - V23: Satisfaction with your life
q4 - V24: Most people can be trusted
q5 - V55: How much freedom of choice and control over own life
q6 - V56: Do you think most people would try to take advantage of you if they got a chance, or would they try to be fair? people can be trusted
q7 - V59: Satisfaction with financial situation of household
q8 - V98: Government responsibility
q9 - V100: Hard work brings success
q10 - V170: Secure in neighborhood
q11 - V237: Family savings during past year
q12 - V238: Social class (subjective)
q13 - V240: Sex
q14 - V242: Age
q15 - V248: Highest educational level attained

```
library(readxl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
survey <- read_excel('WV6.xlsx') # load data
mydata <- select(survey, 11,12,24,25,57,58,62,102,104,201,306,307,309,311,319) # select questions
```

Rename columns and clean no response data

Since for code -1, -2, -3 here and throughout the interview stand for "Don't know", "No answer" and "Not applicable", we decided to remove such elemtents.

```
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------ tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.1     v readr   1.3.1
## v tibble  2.1.3     v purrr   0.3.2
## v tidyr   1.0.0     v stringr 1.4.0
## v ggplot2 3.2.1     v forcats 0.4.0
```

```
## -- Conflicts --------------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
names(mydata)[1] <- "q1"
names(mydata)[2] <- "q2"
names(mydata)[3] <- "q3"
names(mydata)[4] <- "q4"
names(mydata)[5] <- "q5"
names(mydata)[6] <- "q6"
names(mydata)[7] <- "q7"
names(mydata)[8] <- "q8"
names(mydata)[9] <- "q9"
names(mydata)[10] <- "q10"
names(mydata)[11] <- "q11"
names(mydata)[12] <- "q12"
names(mydata)[13] <- "q13"
names(mydata)[14] <- "q14"
names(mydata)[15] <- "q15" # rename column names
mydata[mydata < 0] <- NA # replace negative values with NA
mydata <- drop_na(mydata) # Drop rows containing missing values
```

Parallel Analysis

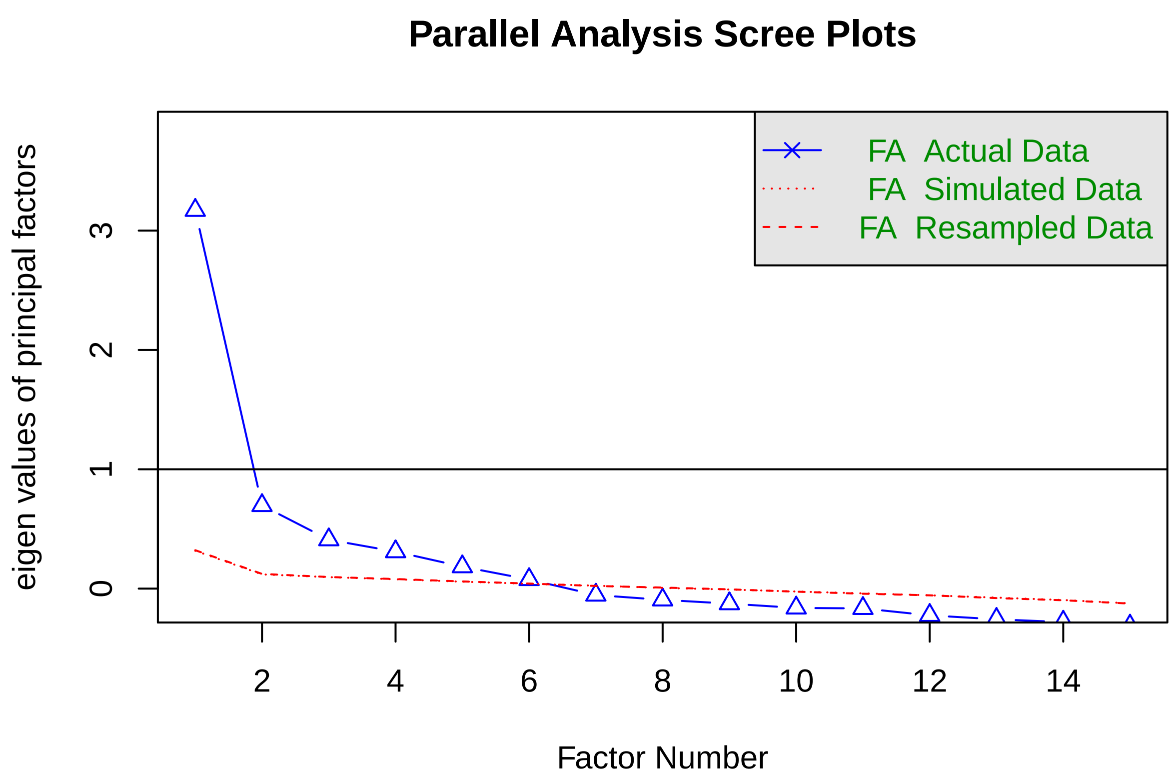
Next we'll find out the number of factors that we'll be selecting for factor analysis.

```r
library(psych)
```

```
##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```r
library(GPArotation)
parallel <- fa.parallel(mydata, fm = 'minres', fa = 'fa') # parallel analysis
```




```
## Parallel analysis suggests that the number of factors =  6  and the number of components =  NA
```

The blue line shows eigenvalues of actual data and the two red lines (placed on top of each other) show simulated and resampled data. Here we look at the large drops in the actual data and spot the point where it levels off to the right. Also we locate the point of inflection – the point where the gap between simulated data and actual data tends to be minimum. Looking at this plot and parallel analysis, anywhere between 2 to 6 factors factors would be good choice.

Factor Analysis

Now that we've arrived at probable number number of factors, let's start off with 4 as the number of factors.

```
fourfactor <- fa(mydata,nfactors = 4,rotate = "oblimin",fm="minres") # 4 factor analysis
print(fourfactor)
```

```
## Factor Analysis using method =  minres
## Call: fa(r = mydata, nfactors = 4, rotate = "oblimin", fm = "minres")
## Standardized loadings (pattern matrix) based upon correlation matrix
##       MR1   MR2   MR3   MR4     h2   u2 com
## q1  -0.73 -0.01  0.00  0.04 0.5230 0.48 1.0
## q2  -0.47  0.14 -0.03  0.36 0.4215 0.58 2.1
## q3   0.83 -0.02  0.03  0.00 0.7253 0.27 1.0
## q4   0.06  0.02 -0.69  0.01 0.4693 0.53 1.0
## q5   0.57 -0.01  0.09  0.09 0.3838 0.62 1.1
## q6   0.22  0.07  0.56  0.08 0.4335 0.57 1.4
## q7   0.36 -0.53  0.02  0.18 0.6325 0.37 2.0
## q8   0.03 -0.31  0.00  0.15 0.1276 0.87 1.5
## q9  -0.13  0.19 -0.03 -0.16 0.1127 0.89 2.8
## q10 -0.10  0.27 -0.24  0.01 0.2250 0.78 2.3
## q11 -0.03  0.57  0.10 -0.07 0.3170 0.68 1.1
## q12  0.03  0.60 -0.09  0.08 0.3936 0.61 1.1
## q13  0.06  0.05  0.06 -0.02 0.0079 0.99 3.1
## q14 -0.09 -0.14  0.16  0.49 0.2979 0.70 1.5
## q15 -0.19 -0.42  0.16 -0.25 0.2541 0.75 2.4
##
##                        MR1  MR2  MR3  MR4
## SS loadings           2.18 1.55 1.06 0.55
## Proportion Var        0.15 0.10 0.07 0.04
## Cumulative Var        0.15 0.25 0.32 0.35
## Proportion Explained  0.41 0.29 0.20 0.10
## Cumulative Proportion 0.41 0.70 0.90 1.00
##
##  With factor correlations of
##       MR1   MR2   MR3   MR4
## MR1  1.00 -0.43  0.34  0.06
## MR2 -0.43  1.00 -0.35 -0.03
## MR3  0.34 -0.35  1.00  0.08
## MR4  0.06 -0.03  0.08  1.00
##
## Mean item complexity =  1.7
## Test of the hypothesis that 4 factors are sufficient.
##
## The degrees of freedom for the null model are  105  and the objective function was  3.06 with Chi Sq
## The degrees of freedom for the model are 51  and the objective function was  0.17
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  0.04
##
## The harmonic number of observations is  2068 with the empirical chi square  328.37  with prob <  3.6
## The total number of observations was  2068  with Likelihood Chi Square =  341.49  with prob <  1.3e-
##
## Tucker Lewis Index of factoring reliability =  0.903
## RMSEA index =  0.053  and the 90 % confidence intervals are  0.047 0.058
```

```
## BIC =  -47.86
## Fit based upon off diagonal values = 0.98
## Measures of factor score adequacy
##                                                    MR1  MR2  MR3   MR4
## Correlation of (regression) scores with factors   0.92 0.85 0.81  0.68
## Multiple R square of scores with factors          0.84 0.73 0.65  0.46
## Minimum correlation of possible factor scores     0.69 0.45 0.30 -0.09
```

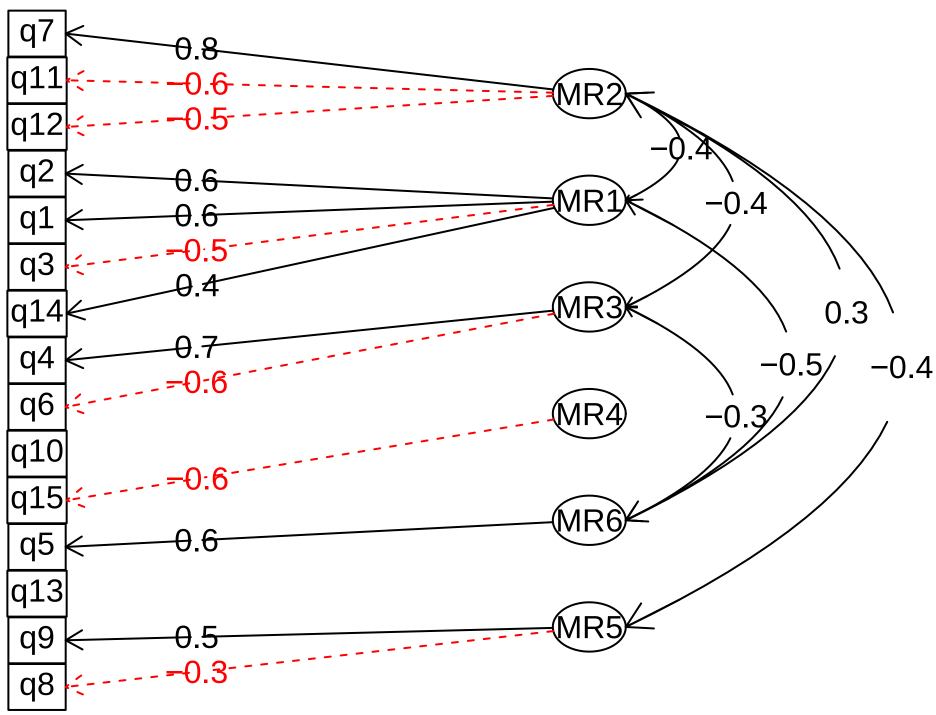
Now we need to consider the loadings more than 0.3 and not loading on more than one factor. Negative values are acceptable here. As you can see 3 variables have become insignificant and 2 other have double-loading. Next, we'll consider '6' factors:

```
sixfactor <- fa(mydata,nfactors = 6,rotate = "oblimin",fm="minres") # 6 factor analysis
print(sixfactor$loadings,cutoff = 0.3)
```

```
##
## Loadings:
##     MR2    MR1    MR3    MR4    MR6    MR5
## q1          0.563
## q2          0.614
## q3         -0.509
## q4                 0.701
## q5                               0.579
## q6                -0.619
## q7   0.773
## q8                                      -0.339
## q9                                       0.495
## q10
## q11 -0.599
## q12 -0.486                0.324
## q13
## q14         0.381
## q15                      -0.627
##
##                  MR2   MR1   MR3   MR4   MR6   MR5
## SS loadings    1.324 1.175 1.001 0.625 0.596 0.536
## Proportion Var 0.088 0.078 0.067 0.042 0.040 0.036
## Cumulative Var 0.088 0.167 0.233 0.275 0.315 0.351
```

```
fa.diagram(sixfactor)
```

Factor Analysis



The root mean square of residuals (RMSR) is 0.01. This is acceptable as this value should be closer to 0. Next we should check RMSEA (root mean square error of approximation) index. Its value, 0.032 shows good model fit as it's below 0.05. Finally, the Tucker-Lewis Index (TLI) is 0.964 – an acceptable value considering it's over 0.9.

Correlations between variables

The first thing to do when conducting a factor analysis or principal components analysis is to look at the correlations of the variables.

```
library(corpcor)
suvMatrix <- cor(mydata) # create a mtrix to show correlations
head(round(suvMatrix, 2))
```

```
##       q1    q2    q3    q4    q5    q6    q7    q8    q9   q10   q11   q12
## q1  1.00  0.41 -0.62  0.14 -0.42 -0.26 -0.41 -0.08  0.14  0.24  0.18  0.19
## q2  0.41  1.00 -0.43  0.12 -0.29 -0.20 -0.32 -0.10  0.10  0.21  0.16  0.22
## q3 -0.62 -0.43  1.00 -0.18  0.53  0.34  0.53  0.15 -0.18 -0.27 -0.21 -0.24
## q4  0.14  0.12 -0.18  1.00 -0.14 -0.43 -0.22 -0.08  0.11  0.26  0.09  0.18
## q5 -0.42 -0.29  0.53 -0.14  1.00  0.29  0.38  0.15 -0.24 -0.21 -0.09 -0.17
## q6 -0.26 -0.20  0.34 -0.43  0.29  1.00  0.34  0.08 -0.10 -0.23 -0.10 -0.18
##      q13   q14   q15
## q1 -0.02 -0.04  0.00
## q2 -0.02  0.15 -0.16
## q3  0.05  0.04  0.05
## q4 -0.02 -0.13 -0.16
## q5  0.09  0.10  0.07
## q6  0.07  0.16  0.06
```

```
cortest.bartlett(mydata) # run Bartlett's test
```

```
## R was not square, finding R from data

## $chisq
## [1] 6302.231
##
## $p.value
## [1] 0
##
## $df
## [1] 105
```

For these data, Bartlett's test is highly significant, and therefore factor analysis is appropriate. Then, we could get the determinant:

```
det(suvMatrix) # get the determinant
```

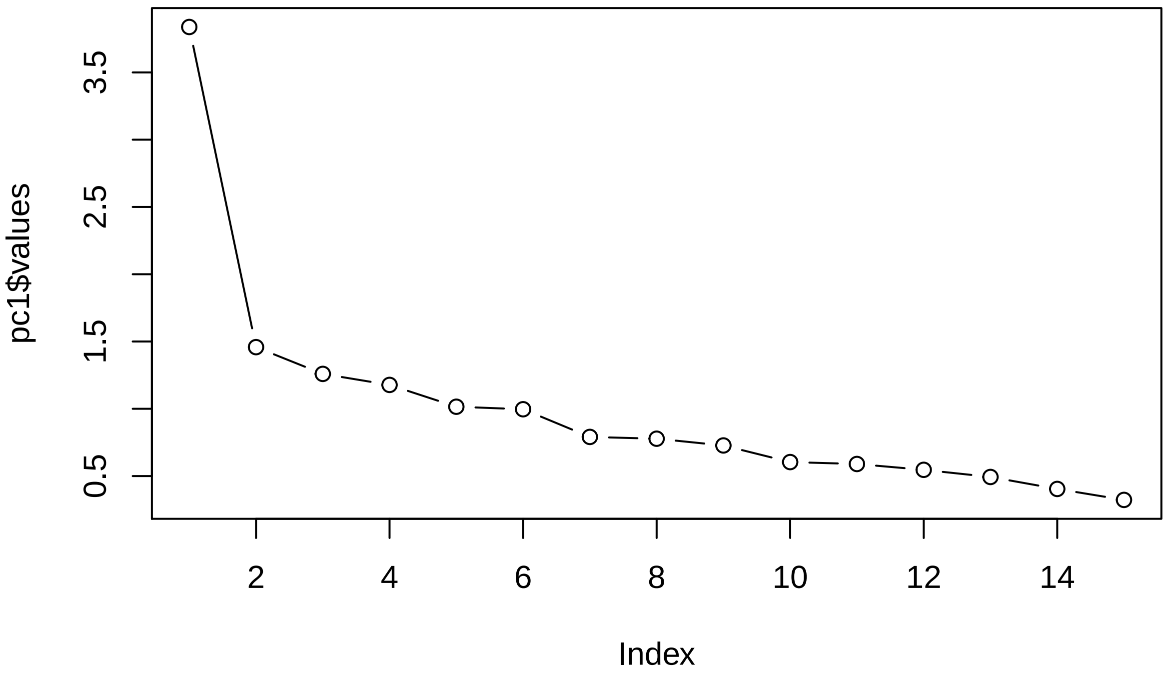
```
## [1] 0.04700017
```

This value is greater than the necessary value of 0.00001. As such, our determinant does not seem problematic.

Factor extraction (PCA)

For our present purposes we will use principal components analysis (PCA).

```
pc1 <- principal(mydata, nfactors=15, rotate="none")
plot(pc1$values, type="b") # scree plot
```



Redo PCA

Now that we know how many components we want to extract, we can rerun the analysis, specifying that number. To do this, we use an identical command to the previous model but we change nfactors = 15 to be nfactors = 6 because we now want only six factors.

```
pc2 <- principal(mydata, nfactors=6, rotate="none")
# factor.model(pc2$loadings)
residuals<-factor.residuals(suvMatrix, pc2$loadings)
residuals<-as.matrix(residuals[upper.tri(residuals)])
```

One approach to looking at residuals is just to say that we want the residuals to be small.

```
large.resid<-abs(residuals) > 0.05
# proportion of the large residuals
sum(large.resid)/nrow(residuals)
```

```
## [1] 0.4761905
```

Rotation

Orthogonal rotation (varimax)

```r
pc3 <- principal(mydata, nfactors=6, rotate="varimax")
print.psych(pc3, cut = 0.3, sort = TRUE) #displaying only loadings above .3
```

```
## Principal Components Analysis
## Call: principal(r = mydata, nfactors = 6, rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##     item   RC1   RC2   RC4   RC6   RC3   RC5   h2    u2 com
## q3     3  0.82                                 0.74 0.262 1.2
## q1     1 -0.80                                 0.66 0.339 1.1
## q5     5  0.67                                 0.58 0.417 1.6
## q2     2 -0.60                     0.46        0.60 0.400 2.1
## q11   11        0.79                           0.66 0.343 1.1
## q12   12        0.70                           0.59 0.410 1.4
## q7     7  0.50 -0.63                           0.72 0.283 2.4
## q4     4             -0.83                     0.69 0.314 1.0
## q6     6  0.33        0.70                     0.64 0.360 1.6
## q10   10             -0.47                     0.42 0.576 3.1
## q9     9                   -0.77               0.63 0.365 1.1
## q8     8                    0.70               0.53 0.470 1.2
## q14   14                           0.68        0.67 0.328 1.9
## q15   15       -0.34              -0.63        0.69 0.305 2.7
## q13   13                                 0.96  0.92 0.082 1.0
##
##                        RC1  RC2  RC4  RC6  RC3  RC5
## SS loadings           2.61 1.84 1.71 1.36 1.18 1.03
## Proportion Var        0.17 0.12 0.11 0.09 0.08 0.07
## Cumulative Var        0.17 0.30 0.41 0.50 0.58 0.65
## Proportion Explained  0.27 0.19 0.18 0.14 0.12 0.11
## Cumulative Proportion 0.27 0.46 0.63 0.77 0.89 1.00
##
## Mean item complexity =  1.6
## Test of the hypothesis that 6 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.08
##  with the empirical chi square  2636.82  with prob <  0
##
## Fit based upon off diagonal values = 0.87
```

Conclusion

According to the results and information of questionnaires above, we could find the questions that load highly on factor 1 are q3 ("Satisfaction with your life") with the highest loading of 0.82, q1 ("Feeling of happiness"), q5 ("How much freedom of choice and control over own life"), q2 ("State of health (subjective)") with the lowest loading of -0.60. All these items seem to relate to physical and mental living conditions. Therefore we might label this factor living environment. Similarly, we might label the factor 2 as financial condition, factor 3 health condition, factor 4 trust condition, factor 5 gender and factor 6 working condition.