Factor Analysis of Global Socioeconomic

Tharindu Kariyawasam - S/18/813

April 2024

1 Introduction

In this study, we are going to perform a factor analysis on socioeconomic indicators across different countries. We are willing to determine number of factors needed to explain the variability of given dataset and determine proportion of variability explain by those factors.

2 Methodology

2.1 Dataset

I use dataset of various socioeconomic indicators across different countries. It includes data on child mortality rates, export and import figures, healthcare expenditures, income levels, inflation rates, life expectancy, fertility rates, and Gross Domestic Product (GDP).

Variable Information

- country - Names of the countries
- child_mort - Death of children under 5 years of age per 1000 live births
- exports - Exports of goods and services per capita, presented as a percentage of the GDP per capita
- health - Total health spending per capita, presented as a percentage of GDP per capita
- imports - Imports of goods and services per capita. Given as a percentage of the GDP per capita
- income - Net income per person
- inflation - The measurement of the annual growth rate of the total GDP
- life_expec - The average number of years a newborn child would live if current mortality patterns remain the same
- total_fer - The number of children that would be born to each woman if current age-fertility rates remain the same
- gdpp - The GDP per capita. Calculated as the total GDP divided by the total population.

2.2 Statistical Methods

In this analysis, I have performed an exploratory factor analysis and a confirmatory factor analysis. Under the exploratory factor analysis, I have used both the principal component method and the maximum likelihood method.

3 Results and discussion

3.1 Exploratory Factor analysis

After scaling the data set first I check whether factor analysis can be apply to this dataset by performing bartlett test and Kaiser-Meyer-Olkin test. Here is the output I get.

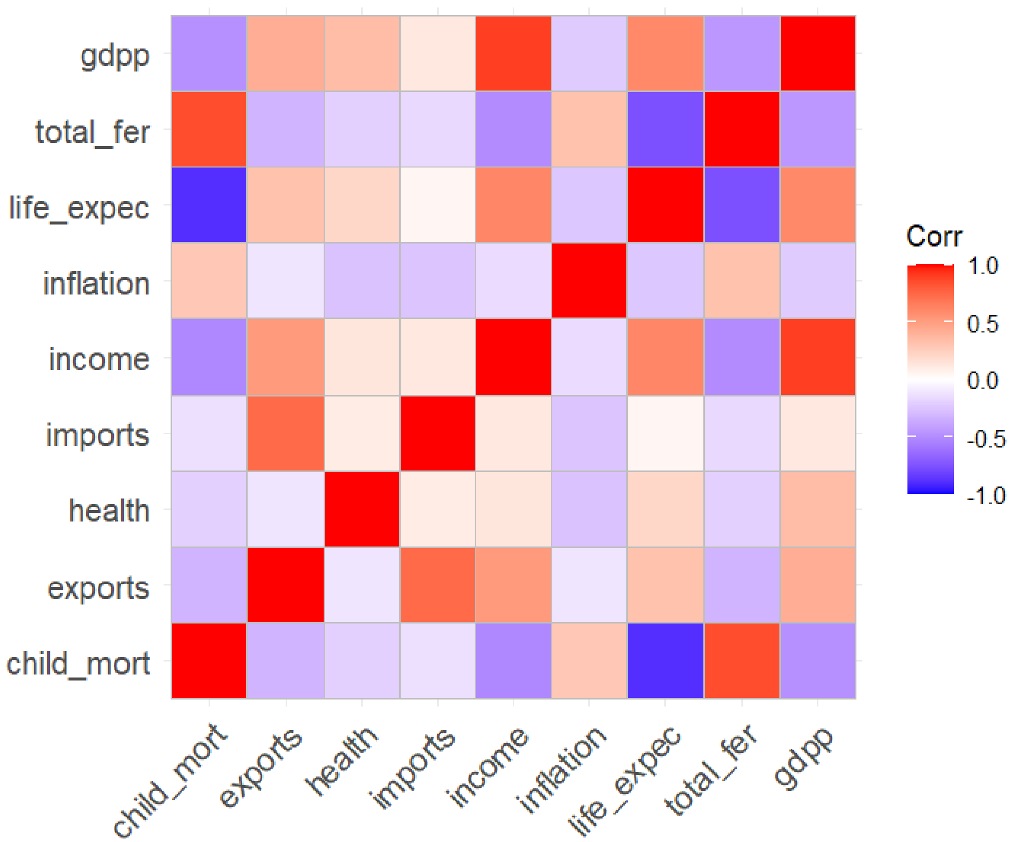
```
> cortest.bartlett(R = cor(Country_Data), n = 167)
$chisq
[1] 1169.737

$p.value
[1] 3.136862e-222

$df
[1] 36
```

```
KMO(Country_Data)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = Country_Data)
Overall MSA =  0.68
MSA for each item =
child_mort    exports     health    imports     income  inflation life_expec  total_fer       gdpp
      0.73       0.57       0.36       0.41       0.69       0.74       0.80       0.86       0.66
```

Since p value of bartlett test is less than 0.05 we can perform factor analysis. And MSA value of KMO test is greater than 0.6 which indicate a medicore value, still we can perform factor analysis.

Then I obtain the correlation matrix and correlogram for the dataset

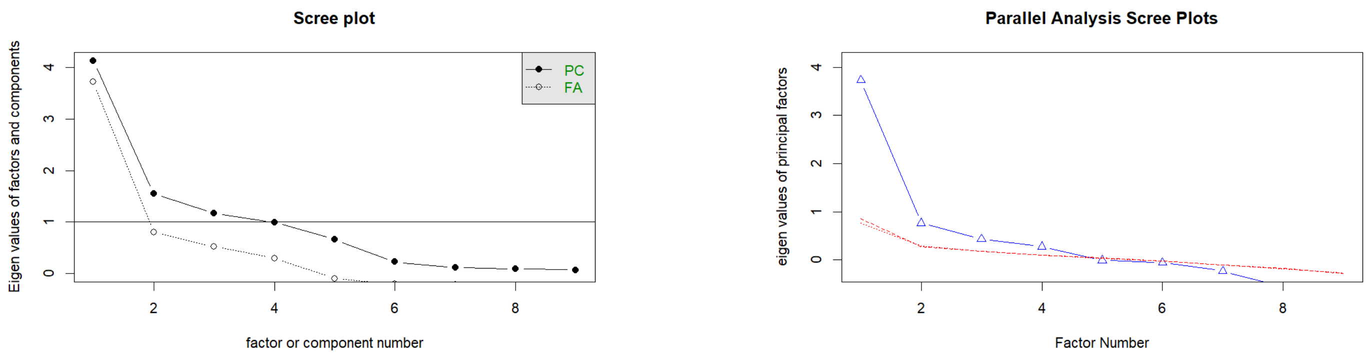
```
> cor.matrix <- cor(Country_Data)
> cor.matrix
           child_mort    exports       health     imports     income  inflation  life_expec  total_fer
child_mort  1.0000000 -0.3180932 -0.20040206 -0.12721092 -0.5243150  0.2882762 -0.88667610  0.8484781 -0.48
exports    -0.3180932  1.0000000 -0.11440840  0.73738083  0.5167836 -0.1072944  0.31631260 -0.3200106  0.41
health     -0.2004021 -0.1144084  1.00000000  0.09571668  0.1295786 -0.2553758  0.21069212 -0.1966740  0.34
imports    -0.1272109  0.7373808  0.09571668  1.00000000  0.1224062 -0.2469943  0.05439053 -0.1590484  0.11
income     -0.5243150  0.5167836  0.12957861  0.12240625  1.0000000 -0.1477560  0.61196247 -0.5018401  0.89
inflation   0.2882762 -0.1072944 -0.25537579 -0.24699428 -0.1477560  1.0000000 -0.23970496  0.3169211 -0.22
life_expec -0.8866761  0.3163126  0.21069212  0.05439053  0.6119625 -0.2397050  1.00000000 -0.7608747  0.60
total_fer   0.8484781 -0.3200106 -0.19667399 -0.15904843 -0.5018401  0.3169211 -0.76087469  1.0000000 -0.45
gdpp       -0.4830322  0.4187248  0.34596553  0.11549817  0.8955714 -0.2216311  0.60008913 -0.4549103  1.00
```



After that I obtained the eigen values and eigen verctors. Then construct scree plot.

```
> e <- eigen(cor.matrix)
> e
eigen() decomposition
$values
[1] 4.13565658 1.54634631 1.17038330 0.99478456 0.66061903 0.22358112 0.11343874 0.08831536 0.06687501

$vectors
            [,1]         [,2]         [,3]         [,4]         [,5]         [,6]         [,7]         [,8]
```

```
[1,]  0.4195194 -0.192883937 -0.02954353  0.370653262  0.16896968 -0.200628153 -0.07948854  0.68274306 -0.
[2,] -0.2838970 -0.613163494  0.14476069  0.003091019 -0.05761584  0.059332832 -0.70730269  0.01419742  0.
[3,] -0.1508378  0.243086779 -0.59663237  0.461897497 -0.51800037 -0.007276456 -0.24983051 -0.07249683 -0.
[4,] -0.1614824 -0.671820644 -0.29992674 -0.071907461 -0.25537642  0.030031537  0.59218953  0.02894642 -0.
[5,] -0.3984411 -0.022535530  0.30154750  0.392159039  0.24714960 -0.160346990  0.09556237 -0.35262369 -0.
[6,]  0.1931729  0.008404473  0.64251951  0.150441762 -0.71486910 -0.066285372  0.10463252  0.01153775  0.
[7,] -0.4258394  0.222706743  0.11391854 -0.203797235 -0.10821980  0.601126516  0.01848639  0.50466425 -0.
[8,]  0.4037290 -0.155233106  0.01954925  0.378303645  0.13526221  0.750688748  0.02882643 -0.29335267  0.
[9,] -0.3926448  0.046022396  0.12297749  0.531994575  0.18016662 -0.016778761  0.24299776  0.24969636  0.
```



(a) Scree Plot

(b) Parallel Analysis Scree Plot

Figure 1: Scree Plots

Then I perform factor analysis for principal component method without rotating and applying varimax rotation. Here is the corresponding factor loadings for both scenarios.

Table 1: Unrotated Factor Loadings

Variable	PA1	PA2	PA3
child_mort	-0.861580023	0.278200112	0.368802431
exports	0.575247653	0.730224600	-0.006013651
health	0.239610472	-0.131399256	0.020524014
imports	0.310156721	0.766087965	-0.252897089
income	0.812619867	0.030704358	0.480372048
inflation	-0.311792649	-0.018169622	0.151838287
life_expec	0.847246566	-0.281776539	-0.142809946
total_fer	-0.788404595	0.196256189	0.313406290
gdpp	0.799958493	-0.029556111	0.524777847

Table 2: Rotated Factor Loadings

Variable	PA1	PA3	PA2
child_mort	0.94651626	-0.22352848	-0.09936125
exports	-0.14607840	0.31677482	0.86167790
health	-0.21535484	0.16259160	-0.04781676
imports	-0.07490014	-0.04021546	0.86012649
income	-0.31829797	0.86825644	0.19201109
inflation	0.30711319	-0.06281341	-0.14945107
life_expec	-0.81132813	0.39672667	0.04443042
total_fer	0.83029791	-0.22226528	-0.13973304
gdpp	-0.30737316	0.89809579	0.12306350

Here is the communality values

Table 3: Communality Estimates

Variable	Communality
child_mort	0.95573067
exports	0.86417399
health	0.07510018
imports	0.74704490
income	0.89205111
inflation	0.12059966
life_expec	0.81761944
total_fer	0.75832180
gdpp	0.91619894

By looking at the communality values we can observe that model explain child_mort, exports, income, life_expec better and imports in a good manner. But for health and inflation model does not explain well.

After that I perform factor analysis for maximum likelihood method as well.

Table 4: Principal Component Method

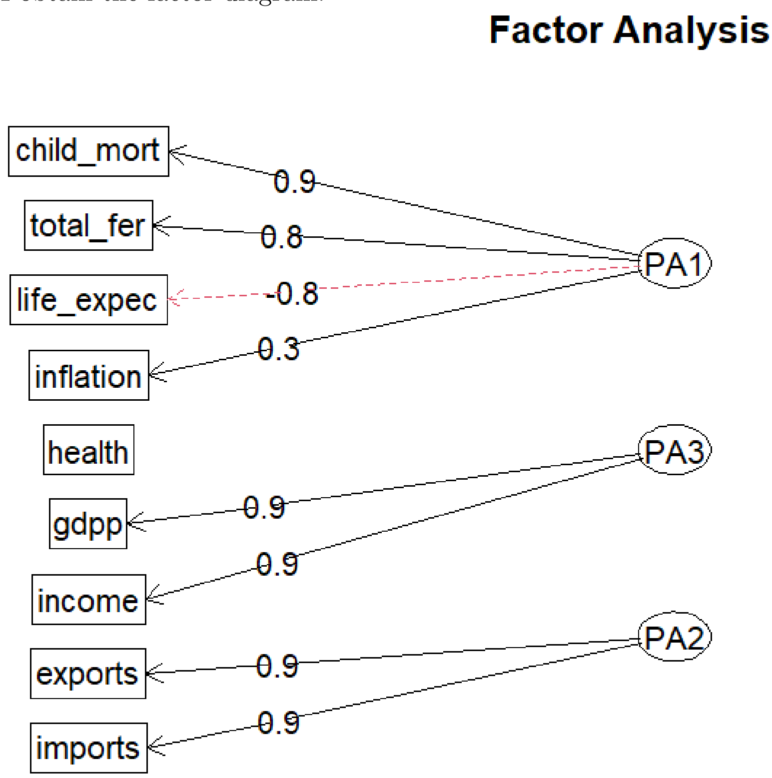
	PA1	PA2	PA3
SS loadings	3.96	1.33	0.85
Proportion Var	0.44	0.15	0.09
Cumulative Var	0.44	0.59	0.68
Proportion Explained	0.64	0.22	0.14
Cumulative Proportion	0.64	0.86	1.00

Table 5: Maximum Likelihood Method

	ML1	ML2	ML3
SS loadings	3.90	1.48	0.90
Proportion Var	0.43	0.16	0.10
Cumulative Var	0.43	0.60	0.70
Proportion Explained	0.62	0.24	0.14
Cumulative Proportion	0.62	0.86	1.00

By above tables we can see that in pricipal component method it explain 0.68 of total variation and in maximum likelihood method it explains 0.70 of total variation.

Then I obtain the factor diagram.



* Hypothesis Test

$$H_0 : \text{Three factors are sufficient}$$
$$H_A : \text{More factors are needed}$$

```
The harmonic n.obs is  167 with the empirical chi square  44.76  with prob <  1.1e-05
```

model probability is less than 0.05. Therefore we reject H_0. We can conclude that the factor model is not sufficient.

3.2 Confirmatory Factor Analysis

```
lavaan 0.6.17 ended normally after 38 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        17

  Number of observations                           167

Model Test User Model:

  Test statistic                               339.586
  Degrees of freedom                                19
  P-value (Chi-square)                           0.000

Model Test Baseline Model:

  Test statistic                              1109.276
  Degrees of freedom                                28
  P-value                                        0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.704
  Tucker-Lewis Index (TLI)                       0.563
```

4 Conclusion and recommendation

* From the analysis, we get three factors to explain the 8 variables of the dataset. None of the three factors have a relationship with variable health.

* Factor 1 is strongly correlated with child_mort, total_fer, life_expec, inflation. We can take that as health factor. Factor 2 is strongly correlated with exports and imports. We can take it as global relations. Factor 3 strongly correlated with gdp and income. We van take it as economic factor.

* 68.3% of the of the total variation is explained by the three factors.

5 References

* Steps of conducting Confirmatory Factor Analysis (CFA) in R "Steps of conducting Confirmatory Factor Analysis (CFA) in R" (n.d.)
* How to run EFA & CFA in R Walker (n.d.)

References

Steps of conducting confirmatory factor analysis (cfa) in r. (n.d.).
Walker, D. L. (n.d.). How to run efa cfa in r.

6 Appendices

6.1 Dataset

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
Afghanista	90.2	10	7.58	44.9	1610	9.44	56.2	5.82	553
Albania	16.6	28	6.55	48.6	9930	4.49	76.3	1.65	4090
Algeria	27.3	38.4	4.17	31.4	12900	16.1	76.5	2.89	4460
Angola	119	62.3	2.85	42.9	5900	22.4	60.1	6.16	3530
Antigua an	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200
Argentina	14.5	18.9	8.1	16	18700	20.9	75.8	2.37	10300
Armenia	18.1	20.8	4.4	45.3	6700	7.77	73.3	1.69	3220
Australia	4.8	19.8	8.73	20.9	41400	1.16	82	1.93	51900
Austria	4.3	51.3	11	47.8	43200	0.873	80.5	1.44	46900

Dataset : `https://www.kaggle.com/datasets/samira1992/countries-intermediate-dataset`

6.2 Code

```
library(tidyverse)
library(psych)
library(lavaan)
library(GPArotation)
library(nFactors)
library(knitr)
library(ggcorrplot)
library(factoextra)

Country_Data <- read.csv(file = "../Data/Country_Data.csv",header = TRUE)
str(Country_Data)
Country_Data <- Country_Data[,-1]

Country_Data <- scale(Country_Data)
colSums(is.na(Country_Data))

cortest.bartlett(R = cor(Country_Data), n = 167)
KMO(Country_Data)
```

```
cor.matrix <- cor(Country_Data)
ggcorrplot(cor.matrix)

e <- eigen(cor.matrix)
scree(Country_Data)
fa.parallel(Country_Data, fa="fa", fm="ml", show.legend=F)

Country_Data_PC<- fa(cor.matrix ,nfactors = 3,rotate = "none",n.obs = 167 ,covar = FALSE,fm = "pa")
Country_Data_ML<- fa(cor.matrix ,nfactors = 3,rotate = "none",n.obs = 167 ,covar = FALSE,fm = "ml")
Rotated_Country_Data_PC<- fa(cor.matrix ,nfactors = 3,rotate = "varimax",n.obs = 167 ,covar = FALSE,fm = "p

fa.diagram(Rotated_Country_Data_PC)

model <- '
  Factor1 =~ child_mort+life_expec+total_fer+inflation
  Factor3 =~ gdpp+income
  Factor2 =~exports+imports
'
variables <- Country_Data[,c("child_mort", "exports", "income", "inflation", "life_expec", "total_fer", "gd

fit <- cfa(model, data = variables)
summary(fit, fit.measures = TRUE)
parameterEstimates(fit, standardized = TRUE, ci = TRUE)
```