the WINNOWER

SOCIAL SCIENCES

Indian states: G and S factors

EMIL O. W. KIRKEGAARD

READ REVIEWS

WRITE A REVIEW

CORRESPONDENCE:
emil@emilkirkegaard.dk

DATE RECEIVED:
June 10, 2015

,
© Kirkegaard This article is distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and redistribution in any medium, provided that the original author and source are credited.

Abstract

I reanalyze data published by Lynn and Yadav (2015) for Indian states. I find both G and S factors which correlate at .61.

The statistical language R is used thruout the paper and the code is explained. The paper thus is both an analysis as a walkthru of how to conduct this type of study.

Introduction

Richard Lynn and Prateek Yadav (2015) have a new paper out in *Intelligence* reporting various cognitive measures, socioeconomic outcomes and environmental factors in some 33 states and areas of India. Their analyses consist entirely of reporting the correlation matrix, but they list the actual data in two tables as well. This means that someone like me can reanalyze it.

They have data for the following variables:

1.
Language Scores Class III (T1). These data consisted of the language scores of class III 11–12 year old school students in the National Achievement Survey (NAS) carried out in Cycle-3 by the National Council of Educational Research and Training (2013). The population sample comprised 104,374 students in 7046 schools across 33 states and union territories (UTs). The sample design for each state and UT involved a three-stage cluster design which used a combination of two probability sampling methods. At the first stage, districts were selected using the probability proportional to size (PPS) sampling principle in which the probability of selecting a particular district depended on the number of class 5 students enrolled in that district. At the second stage, in the chosen districts, the requisite number of schools was selected. PPS principles were again used so that large schools had a higher probability of selection than smaller schools. At the third stage, the required number of students in each school was selected using the simple random sampling (SRS) method. In schools where class 5 had multiple sections, an extra stage of selection was added with one section being sampled at random using SRS.

The language test consisted of reading comprehension and vocabulary, assessed by identifying the word for a picture. The test contained 50 items and the scores were analyzed using both Classical Test Theory (CTT) and Item Response Theory (IRT). The scores were transformed to a scale of 0–500 with a mean of 250 and standard deviation of 50. There were two forms of the test, one in English and the other in Hindi.

2.

Mathematics Scores Class III (T2). These data consisted of the mathematics scores of Class III school students obtained by the same sample as for the Language Scores Class III described above. The test consisted of identifying and using numbers, learning and understanding the values of numbers (including basic operations), measurement, data handling, money, geometry and patterns. The test consisted of 50 multiple-choice items scored from 0 to 500 with a mean score was set at 250 with a standard deviation of 50.

3.

Language Scores Class VIII (T3). These data consisted of the language scores of class VIII (14–15 year olds) obtained in the NAS (National Achievement Survey) a program carried out by the National Council of Educational Research and Training, 2013) Class VIII (Cycle-3).The sampling methodology was the same as that for class III described above. The population sample comprised 188,647 students in 6722 schools across 33 states and union territories. The test was a more difficult version of that for class III, and as for class III, scores were analyzed using both Classical Test Theory (CTT) and Item Response Theory (IRT), and were transformed to a scale of 0–500 with a mean 250.

4.

Mathematics Scores Class VIII (T4). These data consisted of the mathematics scores of Class VIII (14–15 year olds) school students obtained by the same sample as for the Language Scores Class VIII described above. As with the other tests, the scores were transformed to a scale of 0–500 with a mean 250 and standard deviation of 50.

5.

Science Scores Class VIII (T5). These data consisted of the science scores of Class VIII (14–15 year olds) school students obtained by the same sample as for the Language Scores Class VIII described above. As with the other tests, the scores were transformed to a scale of 0–500 with a mean 250 and standard deviation of 50. The data were obtained in 2012.

6.

Teachers' Index (TI). This index measures the quality of the teachers and was taken from the Elementary State Education Report compiled by the District Information System for Education (DISE, 2013). The data were recorded in September 2012 for teachers of grades 1–8 in 35 states and union territories. The sample consisted of 1,431,702 schools recording observations from 199.71 million students and 7.35 million teachers. The teachers' Index is constructed from the percentages of schools with a pupil–teacher ratio in primary greater than 35, and the percentages single-teacher schools, teachers without professional qualification, and female teachers (in schools with 2 and more teachers).

7.

Infrastructure Index (II). These data were taken from the Elementary State Education Report 2012–13 compiled by the District Information System for Education (2013). The sample was the same as for the Teachers' Index described above. This index measures the infrastructure for education and was constructed from the percentages of schools with proper chairs and desks, drinking water, toilets for boys and girls, and with kitchens.

8.

GDP per capita (GDP per cap). These data are the net state domestic product of the Indian states in 2008–09 at constant prices given by the Reserve Bank of India (2013). Data are not available for the Union Territories.

9.

Literacy Rate (LR). This consists of the percentage of population aged 7 and above in given in the 2011 census published by the Registrar General and Census Commission of India (2011).

10.

Infant Mortality Rate (IMR). This consists of the number of deaths of infants less than one year of age per 1000 live births in 2005–06 given in the National Family Health Survey, Infant and Child Mortality given by the Indian Institute of Population Sciences (2006).

11.

Child Mortality Rate (CMR). This consists of the number of deaths of children 1–4 years of age per 1000 live births in the 2005–06 given by the Indian Institute of Population Sciences (2006).

12.

Life Expectancy (LE). This consists of the number of years an individual is expected to live after birth, given in a 2007 survey carried out by Population Foundation of India (2008).

13.

Fertility Rate (FR). This consists of the number of children born per woman in each state and union territories in 2012 given by Registrar General and Census Commission of India (2012).

14.

Latitude (LAT). This consists of the latitude of the center of the state.

15.

Coast Line (CL). This consists of whether states have a coast line or are landlocked and is included to examine whether the possession of a coastline is related to the state IQs.

16.

Percentage of Muslims (MS). This is included to examine a possible relation to the state IQs.

This article will include the R code line for line commented as a helping exercise for readers not familiar with R but who can perhaps be convinced to give it a chance! :)

```
library(devtools) #source_url
source_url("https://osf.io/j5nra/?action=download&version=2") #mega functions from OSF
#source("mega_functions.R")
library(psych) #various
library(car) #scatterplot
library(Hmisc) #rcorr
library(VIM) #imputation
```

This loads a variety of libraries that are useful.

Getting the data into R

```
cog = read.csv("Lynn_table1.csv",skip=2,header=TRUE,row.names = 1) #load cog data
socio = read.csv("Lynn_table2.csv",skip=2,header=TRUE,row.names = 1) #load socio data
```

The files are the two files one can download from ScienceDirect: Lynn_table1 Lynn_table2 The code makes it read it assuming values are divided by comma (CSV = comma-separated values), skips the first two lines because they do not contain data, loads the first line as headers, and uses the first column as rownames.

Merging data into one object

Ideally, I'd like all the data as one object for easier use. However, since it comes it two, it has to be merged. For this purpose, I rely upon a dataset merger function I wrote some months ago to handle international data. It can however handle any merging of data where one wants to match rows by name from different datasets and combine them into one dataset. This function, merge_datasets(), is found in the mega_functions we imported earlier.

However, first, it is a good idea to make sure the names do match when they are supposed to. To check this we can type:

```
cbind(rownames(cog),rownames(socio))
```

I put the output into Excel to check for mismatches:

Andhra Pradesh	Andhra Pradesh	TRUE
Arunachal Pradesh	Arunachal Pradesh	TRUE

```
Bihar            Bihar            TRUE
Chattisgarh      Chattisgarh      TRUE
Goa              Goa              TRUE
Gujarat          Gujarat          TRUE
Haryana          Haryana          TRUE
Himanchal PradeshHimanchal PradeshTRUE
Jammu Kashmir    Jammu & Kashmir  FALSE
Jharkhand        Jharkhand        TRUE
Karnataka        Karnataka        TRUE
Kerala           Kerala           TRUE
Madhya Pradesh   Madhya Pradesh   TRUE
Maharashtra      Maharashtra      TRUE
Manipur          Manipur          TRUE
Meghalaya        Meghalaya        TRUE
Mizoram          Mizoram          TRUE
Nagaland         Nagaland         TRUE
Odisha           Odisha           TRUE
Punjab           Punjab           TRUE
Rajashthan       Rajasthan        FALSE
Sikkim           Sikkim           TRUE
Tamil Nadu       TamilNadu        FALSE
Tripura          Tripura          TRUE
Uttarkhand       Uttarkhand       TRUE
Uttar Pradesh    Uttar Pradesh    TRUE
West Bengal      West Bengal      TRUE
A & N Islands    A & N Islands    TRUE
Chandigarh       Chandigarh       TRUE
D & N Haveli     D & N Haveli     TRUE
Daman & Diu      Daman & Diu      TRUE
Delhi            Delhi            TRUE
Puducherry       Puducherry       TRUE
```

So we see that the order is the same, however, we see that there are three that doesn’t match despite being supposed to. We can fix this discrepancy by using the rownames of one dataset for the other:

```
rownames(cog) = rownames(socio) #use rownames from socio for cog
```

This makes the rownames of cog the same as those for socio. Now they are ready for merging.

Incidentally, since the order is the same, we could have simply merged with the command:

```
cbind(cog, socio)
```

However it is good to use merge_datasets() since it is so much more generally useful.

Missing and broken data

Next up, we examine missing data and broken data.

```
#examine missing data
miss.table(socio)
miss.table(cog)
table(miss.case(socio))
matrixplot(socio)
```

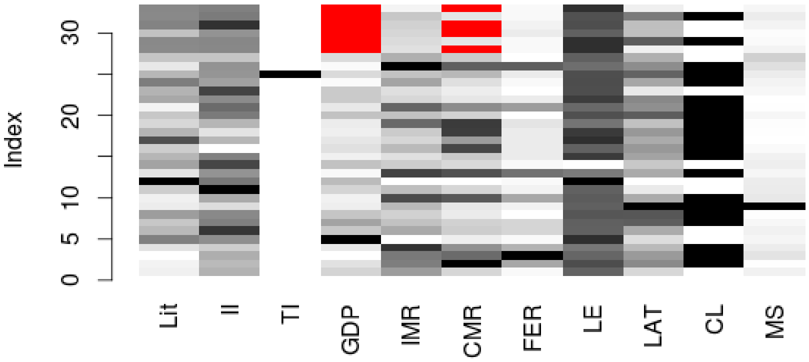
The first, miss.table(), is another custom function from mega_functions. It outputs the number of missing values per variable. The outputs are:

```
Lit  II  TI GDP IMR CMR FER  LE LAT  CL  MS
  0   0   0   6   0   4   0   0   0   0   0
T1 T2 T3 T4 T5 CA
 0  0  0  0  0  0
```

So we see that there are 10 missing values in the socio and 0 in cog.

Next we want to see how these are missing. We can do this e.g. by plotting it with a nice function like matrixplot() (from VIM) or by tabling the missing cases. Output:

```
 0  1  2
27  2  4
```



So we see that there are a few cases that miss data from 1 or 2 variables. Nothing serious.

One could simply ignore this, but that would be not utilizing the data to the full extent possible. The correct solution is to impute data rather than removing cases with missing data.

However, before we do this, look at the TI variable above. The greyscale shows the standardized values of the datapoints. So in this variable we see that there is one very strong outlier. If we take a look back at the data table, we see that it is likely an input error. All the other datapoints have values between 0 and 1, but the one for Uttarkhand has 247,200.595… I don't see how the input error happened the so best way is to remove it:

```
#fix broken datapoint
socio["Uttarkhand","TI"] = NA
```

Then, we impute the missing data in the socio variable:

```
#impute data
socio2 = irmi(socio, noise.factor = 0) #no noise
```

The second parameter is used for multiple imputation, which we don't use here. Setting it as 0 means that the imputation is deterministic and hence exactly reproducible for other researchers.

Finally, we can compare the non-imputed dataset to the imputed one:

```
#compare desp stats
describe(socio)
describe(socio2)
round(describe(socio)-describe(socio2),2) #discrepancy values, rounded
```

The output is large, so I won't show it here, but it shows that the means, sd, range, etc. of the variables with and without imputation are similar which means that we didn't completely mess up the data by the procedure.

Finally, we merge the data to one dataset:

```
#merge data
data = merge.datasets(cog,socio2,1) # merge above
```

Next, we want to do factor analysis to extract the general socioeconomic factor and the general intelligence factor from their respective indicators. And then we add them back to the main dataset:

```
#factor analysis
fa = fa(data[1:5]) #fa on cognitive data
```

```
fa
data["G"] = as.numeric(fa$scores)

fa2 = fa(data[7:14]) #fa on SE data
fa2
data["S"] = as.numeric(fa2$scores)
```

Columns 1-5 are the 5 cognitive measures. Cols 7:14 are the socioeconomic ones. One can disagree about the illiteracy variable, which could be taken as belonging to cognitive variables, not the socioeconomic ones. It is similar to the third cognitive variable which is some language test. I follow the practice of the authors.

The output from the first factor analysis is:

```
    MR1     h2   u2 com
T1 0.40 0.1568 0.84   1
T2 0.10 0.0096 0.99   1
T3 0.46 0.2077 0.79   1
T4 0.93 0.8621 0.14   1
T5 0.92 0.8399 0.16   1
Proportion Var 0.42
```

This is using the default settings, which is minimum residuals. Since the method used typically does not matter except for PCA on small datasets, this is fine.

All loadings are positive as expected, but T2 is only slightly so.

We put the factor scores back into the dataset and call it “G” (Rindermann, 2007).

The factor analysis output for socioeconomic variables is:

```
      MR1    h2   u2 com
Lit  0.79 0.617 0.38   1
II   0.36 0.128 0.87   1
TI   0.91 0.824 0.18   1
GDP  0.76 0.579 0.42   1
IMR -0.92 0.842 0.16   1
CMR -0.85 0.721 0.28   1
FER -0.84 0.709 0.29   1
LE   0.14 0.019 0.98   1
Proportion Var 0.55
```

Strong positive loadings for: proportion of population literate (LIT), teacher index (TI), GDP, medium positive for infrastructure index (II), weak positive for life expectancy (LE). Strong negative for infant mortality rate (IMR), child mortality rate (CMR) and fertility. All of these are in the expected direction.

Then we extract the factor scores and add them back to the dataset and call them “S”.

Correlations

Finally, we want to check out the correlations with G and S.

```
#Pearson results
results = rcorr2(data)
View(results$r)  #view all correlations
results$r[,18:19] #S and G correlations
results$n #sample size

#Spearman
results.s = rcorr2(data, type="spearman") #spearman
View(results.s$r) #view all correlations

#discrepancies
results.c = results$r-results.s$r
```

We look at both the Pearson and Spearman correlations because data may not be normal and may have outliers. Spearman's is resistant to these problems. The discrepancy values are how larger the Pearson is than the Spearman.

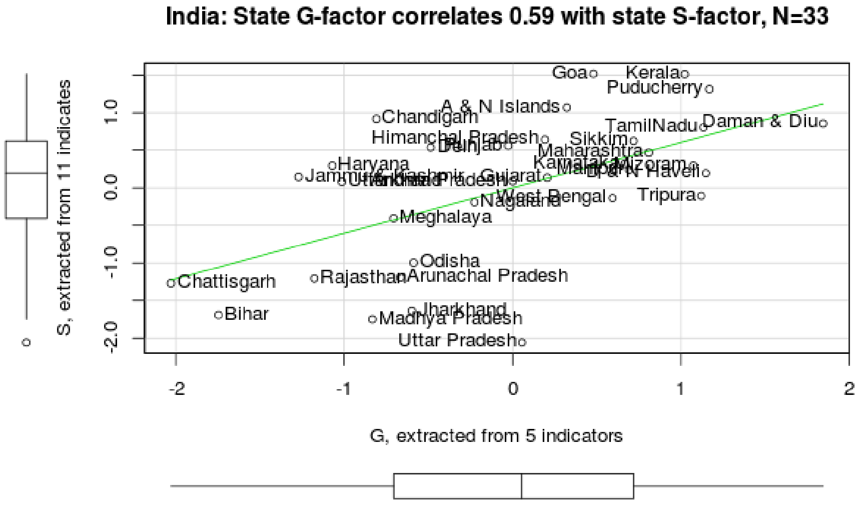
There are too many correlations to output here, so we focus on those involving G and S (columns 18:19).

Variable	G	S
T1	0.41	0.41
T2	0.10	-0.39
T3	0.48	0.16
T4	0.97	0.62
T5	0.96	0.53
CA	0.87	0.38
Lit	0.66	0.81
II	0.45	0.37
TI	0.40	0.93
GDP	0.40	0.78
IMR	-0.60	-0.94
CMR	-0.54	-0.87
FER	-0.56	-0.86
LE	0.01	0.14
LAT	-0.53	-0.34
CL	-0.63	-0.54
MS	-0.24	-0.08
G	1.00	0.59
S	0.59	1.00

So we see that G and S correlate at .59, fairly high and similar to previous within country results with immigrant groups (.54 in Denmark, .59 in Norway Kirkegaard (2014a), Kirkegaard and Fuerst (2014)) but not quite as high as the between country results (.86-.87 Kirkegaard (2014b)). Lynn and Yadav mention that data exists for France, Britain and the US. These can serve for reanalysis with respect to S factors at the regional/state level.

Finally, we want may to plot the main result:

```
#Plots
title = paste0("India: State G-factor correlates ",round(results$r["S","G"],2)," with state S-factor, N=",results$n["S","G"])
scatterplot(S ~ G, data, smoother=FALSE, id.n=nrow(data),
        xlab = "G, extracted from 5 indicators",
        ylab = "S, extracted from 11 indicates",
        main = title)
```



It would be interesting if one could obtain genomic admixture measures for each state and see how they relate, since this has been found repeatedly elsewhere and is a strong prediction from genetic theory.

Update

Lynn has sent me the correct datapoint. It is 0.595. The imputed value was around .72. I reran the analysis with this value and imputed the rest. It doesn't change much. The new results are slightly stronger.

	New results		Discrepancy scores	
	G	S	G	S
T1	0.41	0.42	0.00	-0.01
T2	0.10	-0.37	0.00	-0.02
T3	0.48	0.18	0.00	-0.02
T4	0.97	0.63	0.00	-0.02
T5	0.96	0.54	0.00	-0.01
CA	0.87	0.40	0.00	-0.02
Lit	0.66	0.81	0.00	-0.01
II	0.45	0.37	0.00	0.00
TI	0.42	0.92	-0.02	0.01
GDP	0.40	0.78	0.00	0.00
IMR	-0.60	-0.95	0.00	0.00
CMR	-0.54	-0.87	0.00	0.00
FER	-0.56	-0.86	0.00	-0.01
LE	0.01	0.14	0.00	0.00
LAT	-0.53	-0.35	0.00	0.01
CL	-0.63	-0.54	0.00	0.00
MS	-0.24	-0.09	0.00	0.00
G	1.00	0.61	0.00	-0.01
S	0.61	1.00	-0.01	0.00

Method of correlated vectors

This study is special in that we have two latent variables each with its own set of indicator variables. This means that we can use Jensen's method of correlated vectors (MCV; Jensen (1998)), and also a new version which I shall creatively dub "double MCV", DMCV using both latent factors instead of only one.

The method consists of correlating the factor loadings of a set of indicator variables for a factor with the correlations of each indicator variable with a criteria variable. Jensen used this with the general intelligence factor (g-factor) and its subtests with criteria variables such as inbreeding depression in IQ scores and brain size.

So, to do regular MCV in this study, we first choose either the S and G factor. Then we correlate the loadings of each indicator with its correlation with the criteria variable, i.e. the S/G factor we didn't choose.

Doing this analysis is in fact very easy here, because the results reported in the table above with S and G is exactly that which we need to correlate.

```
## MCV
#Double MCV
round(cor(results$r[1:14,18],results$r[1:14,19]),2)
#MCV on G
round(cor(results$r[1:5,18],results$r[1:5,19]),2)
#MCV on S
round(cor(results$r[7:14,18],results$r[7:14,19]),2)
```

The results are: .87, .89, and .97. In other words, MCV gives a strong indication that it is the latent traits that are responsible for the observed correlations.

References

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.

Kirkegaard, E. O. W. (2014a). Crime, income, educational attainment and employment among immigrant groups in Norway\n and Finland. Open Differential Psychology.

Kirkegaard, E. O. W., & Fuerst, J. (2014). Educational attainment, income, use of social benefits, crime rate and the general so\ncioeconomic factor among 71 immigrant groups in Denmark. Open Differential Psychology.

Kirkegaard, E. O. W. (2014b). The international general socioeconomic factor: Factor analyzing international rankin\ngs. Open Differential Psychology.

Lynn, R., & Yadav, P. (2015). Differences in cognitive ability, per capita income, infant mortality, fertility and l\natitude across the states of India. *Intelligence*, *49*, 179-185.

Rindermann, H. (2007). The g-factor of international cognitive ability comparisons: The homogeneity of result\ns in PISA, TIMSS, PIRLS and IQ-tests across nations. *European Journal of Personality*, *21*(5), 667-706.