WHAT MAKES A COUNTRY GREAT

by Bingjie Shen, Haiqiong Li, Waishing Wong, Yang Xiao (in alphabetical order)

1.Introduction

Our project------What makes a country great is inspired by someone who is keep saying MAGG, we want to explore the relationship that between national variables and how it simply contributes to classify counties and to make a country better than others. Our project will be five parts:

1) PCA method to get the PCs and, also reduce dimension
2) Multivariate PCA regression and hypothesis test
3) Cluster
4) FA and score ranking
5) Conclusion and limitation

2. Data collection and pretreatment

Our data is collected from https://www.worldbank.org/. We deleted some countries out of our list since they have a lot of missing data. And make up few missing data by filling with mean value. Because national data is collected under different unit, before doing PCA, we standardize all the variables so that they are under the same unit. The raw data which contain 44 variables can be reached from attachment, also the data index.

3.PCA

3.1 Objective and method introduction

3.1.1 Objective

By applying PCA method, variable dimension will be reduced, and we will decide the number of PCs need to be retained to explain the most of information of raw data. Also, how dose each variable contributes to PCs.

3.1.2 method introduction

PCA is a dimensionality reduction technique. It is a method that groups data into principle components (PC) based on variances, where PC1 has the highest variance along an axes, PC 2 has the second highest...etc. Each PC is a linear combination of the variables, such as PC1 = A*x1 + B*x2 +.... C*x43.

PCA is usually used for few functionalities. First, it is used for visualization. PCA allows for visualizations of data relationships in high dimensional space, such as plotting PC1 vs. PC2. For example, clusters in a PC1 vs. PC2 graph are likely similar data points. Alternatively, we may see how different the data are given the distance among them in a PC1 vs. PC2 graph. (Recall that axes are ranked in order of % variance explained. Hence, the differences along PC1 are greater than the differences along PC2.)

Second, it is used to assist a regression or a classification procedure. If a data has many variables, it may cause researchers to over-fit if he does not apply techniques such as regularizations or dimensionality reduction. In PCA, each component is independent of one another, thus eliminating the possibility of over-fitting by including correlated variables. In this report, the objective of PCA is to reduce the dimensionality for variables x3 to x43, get an output of PCs to regress against x44(Urban population growth).

3.1.3 PCA assumptions & limitations

- The variables must be continuous
- There need to be a linear relationship among all variables
- There should be a sampling adequacy. Sample size must be sufficiently large for PCA to work properly.
- There should be no significant outliers. PCA groups the variables into PCs based on variance. Hence, a large outlier will have influence on the PCA results.

3.2 Result and interpretation

There are multiple methods to decide number of principle components to keep. We shall proceed with two of them. For the purpose of this project, the threshold is 80% variance. According to the results, method 1 suggests us to keep the first 13 principle components.

The 12th eigenvalue is 1.08, and the 13th eigenvalue is 0.86. According to the results, the average value is 1, which suggests we shall keep the first 12 principle components (More detail about variance explained, see figure 3.1 in appendix).

Since the two methods show different results, it is up to the researchers to decide the proper number of PCs to keep based on the 2 methods. Notice that 12 PCs explained about 79.84% of the variance, while 13 PCs explained 82.01%. This is not a significantly huge increase in % variance explained, and 79.94% is fairly close to the 80% threshold. Due to the reasons mentioned above, the first 12 PCs shall be kept.

As a result of the PCA method, 12 PCs are received as an output. They may be used in the regression procedure in part 2. The complete list of coefficients in the 12 PCs is attached as 'PCA.xlsx'. In the appendix, a snippet of the coefficients for PC1 is shown in graphical view as an example.

3.3 Summary

Method 1 is to decide number of components to keep based on some determined threshold. This threshold shall limit the possibility of keeping too many principle components, while ensuring the PCs explained a certain amount of variance.

Method 2 is to drop PCs that are under the average eigenvalue. Since our data is standardized, so the average eigenvalue is 1, we can just choose PCs than whose eigenvalue is larger than 1.

As a result of the PCA method, 12 PCs are received as an output. They may be used in the regression procedure in part 4. The complete list of coefficients in the 12 PCs is attached as 'PCA.xlsx'.

4.Multivariate regression

4.1 Objective and method introduction

4.1.1 Objective

In this section, we will use multivariate regression method to determine which variables are linearly related with GDP and Urban population growth rate.

4.1.2 Method introduction

Multivariate Regression is a method used to measure the degree at which more than one independent variable (predictors) and more than one dependent variable (responses), are linearly related. By using this method, we want to know which variables are most linearly related with GDP and Urban population growth rate. For example, we want to determine which variables increase, the GDP and Urban population growth rate will also increase.

4.1.3 Assumption

- E(**Y**) = **XB**.
- Cov (yi) = 0 for all i = 1,2,3, n.
- Cov (yi; yj) = 0 for all i 6= j.

First, means of response variables are equal to linear formula between predictive variables and parameter. Second, response variables **Y** has correlated columns but uncorrelated rows.

4.2 Result and interpretation

As we can see from the Table4.2(see appendix), the top 4 variables with the highest coefficient values relate GDP are x19(GNI), X20(high –technology exports), x25(Land area) and X35(Population amount). Respectively, these four variables' coefficient rate related to GDP are 0.286, 0.274, 0.201 and 0.175. Thus, 1unit growth in gross national income, the GDP will increase in 0.286. And, 1 unit increase in high-technology exports, the GDP will increase in 0.274. Also, the GDP will increase in 0.201, if there is 1 unit increase in land area. Last but not least, if the total population increase in 1 unit, the GDP will increase in 0.175.

When related to the urban population growth rate, the top 4 variables with the highest absolute coefficient values are X7 (death rate), X14(forest area), X24(labor force partition rate), and X34(population growth rate). Respectively, these four variables' coefficient values are -.1998, -.1523, 0.1218, 0.2598. Thus, 1 unit increase in the death rate, the urban population growth rate will decrease by 0.1998. And, 1 unit increase in the forest area, the urban population growth rate will decrease in 0.1523. Also, the urban population growth rate will increase by 0.1218, if there is 1 unit increase in the labor force partition rate. Last but not least, if the population growth rate increases in 1 unit, the urban population growth rate will increase in 0.2598.

4.3 Summary

Based on the analysis above, as the governors, if they want to increase the GDP, they may put more efforts in increasing the gross national income, high-technology exports, and total population amount. Besides, if the labor force partition rate and the population increasing rate are increasing, the governors may anticipate that the urban population growth rate will also increase so that to get prepared for the increasing urban population.

5.Cluster

5.1 Objective and method introduction

5.1.1 Objective

The goal for cluster analysis is to search for patterns in this data set by grouping observations into clusters. In our case, we would like to see if there is any pattern between all the countries with 41features.

5.1.2 Method introduction

In cluster analysis, the decision of merging two clusters is taken on the base of closeness of these clusters. There are multiple ways to measure the distance for deciding the closeness of two clusters:

- Euclidean distance: d (**x**, **y**) = $\sqrt{(\boldsymbol{x}-\boldsymbol{y})'(\boldsymbol{x}-\boldsymbol{y})} = \sqrt{\sum_{j=1}^{p}\left(x_j - y_j\right)^2}$.
- Mahalanobis distance: d (**x**, **y**) = $\sqrt{(\boldsymbol{x}-\boldsymbol{y})'\boldsymbol{S}^{-1}(\boldsymbol{x}-\boldsymbol{y})}$, where **S** is the sample covariance matrix.
- Minkowski distance: d (**x**, **y**) =$\left[\sum_{j=1}^{p}\left|x_j - y_j\right|^r\right]^{\frac{1}{r}}$.

Two common methods of cluster analysis:

Hierarchical clustering

Within this approach to cluster analysis there are numbers of different methods used to determine which clusters should be joined at each stage. The main methods are summarized below:

- Single linkage method (Nearest neighbor method)
- Average linkage method
- Complete linkage method (Farthest neighbor)
- Centroid method

The single linkage method is relatively simple, but it is sensitive to errors in distances between observations. The complete linkage method tends to produce compact clusters of similar size,

but does not take account of cluster structure. It is quite sensitive to outlier. Both Average link method and Centroid method are fairly robust.

K-means clustering method

In this method, the desired number of clusters is specified in advance and the 'best' solution is chosen. In general, it follows those five as follows:

- Choose initial cluster centers (essentially this is a set of observations that are far apart — each subject forms a cluster of one and its center is the value of the variables for that subject).
- Assign each subject to its 'nearest' cluster, defined in terms of the distance to the centroid.
- Find the centroids of the clusters that have been formed
- Re-calculate the distance from each subject to each centroid and move observations that are not in the cluster that they are closest to.
- Continue until the centroids remain relatively stable.

5.1.3 Limitations

Cluster analysis does not make any distinction between dependent and independent variables, which is unlike Factor analysis. Therefore, the clusters formed can be very dependent on the variables included.

5.2 Result and interpretation

5.2.1 Hierarchical clustering to Determine how many clusters

We use the Hierarchical clustering (average link method) approach initially to determine how many clusters there are in the data. We used two methods to choose g.

Method 1: select g clusters from the dendrogram by cutting across the branches at a given level of the distance measure used by one of the axes.

Based on the SAS output, it indicates the largest change in levels occurs in going from 7 clusters to 8 cluster. The change in root squared distance between the 7-cluster solution and the 8-cluster solution is 2.103. The difference between the 8-cluster solution and the 9-cluster solution is 0.294. Therefore, we chose 7 clusters. (Figure 5.1 in appendix)

Method 2: $\alpha_j > \overline{\alpha} + ks_\alpha$

Given k=1.25, $\alpha_j > \overline{\alpha} + ks_\alpha$ = 2.807+3.451*2.75=12.297. a_6=12.735> 12.297and a_7<12.297, so we choose 6 clusters.

Based on those two methods discussed above, the initial clusters can be 7 or 6.

5.2.2 Conduct K-means clustering

For K-means clustering, we did both random 6 observations and first 7 observations. The idea is to see which approach would give us a better and more reliable clustering.

Random 6 observations (Figure 5.2 in appendix)

As the cluster summary table given by SAS, we can see cluster 3 has the largest number of countries, which includes 37 countries, while cluster 2 has the least number of countries, which only includes 2 countries. As we can see from the figure 5.2, those two countries are China and United States. Based on the plots of discriminant functions, we can see all the countries are well divided into two groups instead of six groups. Additionally, only cluster 4 and cluster 5 are sufficiently separated from other clusters.

First 7 observations (Figure 5.3 and 5.4 in appendix)

As the cluster summary table given by SAS, we can see cluster 5 has the largest number of countries, which includes 41 countries, while cluster 2 has the least number of countries, which only includes 2 countries. As we can see from the figure 5.4, those two countries are China and India. Based on the plots of discriminant functions, we can see all the countries are well divided into four groups instead of six groups. Additionally, we can see the first 7 observation method gives us a better clustering result. As we can see from the plot below, cluster 4, cluster 6 and cluster 2 are well separated from other clusters. Within each group, all the data are more compacted.

5.3 Summary

Let us focus on the more compacted outcome which is shown in figure 5.4. There are two clusters that just has few countries, one cluster has China and India who are two largest developed countries, having the largest number of population in the world, the other cluster has Russian, the USA, Canada, Brazil and Australia, most of this cluster are developed countries that have generally more advanced post-industrial economies, meaning the service sector provides more wealth than the industrial sector. That outcomes are fairly accord with the real-world's common sense, which means our cluster is meaningful, to some degree.

6.FA and ranking

6.1 Objective and method introduction

6.1.1 Objective

We are planning to use factor analysis method to get main factors of all 41 variables to score each factor, and then apply weighted eigenvalue with each factor after rotation to come with a total score of country, so we can do ranking and interpretation base on ranking in the end.

6.1.2 Method introduction

The factors are underlying constructs or latent variables that "generate" the y's. Like the original variables, the factors vary from individual to individual; but unlike the variables, the factors cannot be measured or observed. In our report, we used principal components method to get the factors. factor scores, fˆi = (fˆi1, fˆi2, . . . , fˆim)", i = 1, 2, . . . , n, which are defined as estimates of the underlying factor values for each observation. There are two potential uses for such scores: (1) the behavior of the observations in terms of the factors may be of interest and (2) we may wish to use the factor scores as input to another analysis, such as MANOVA.

6.1.3 Limitation

- Its usefulness depends on the researchers' ability to develop a complete and accurate set of product attributes - If important attributes are missed the value of the procedure is reduced accordingly.
- Naming of the factors can be difficult - multiple attributes can be highly correlated with no apparent reason.
- If the observed variables are completely unrelated, factor analysis is unable to produce a meaningful pattern

6.2 Result and interpretation

6.2.1 FA

After choosing factors by using PCA method, we get the same number of factors as the number of PCs that we got before, which is 12, and each factor is a combination of all 41 variables, coefficient in front of each variable mean how much it contributes to factor . Figure 6.1 show us the variance that explained by 12 factors after rotation.

Since we used PCA method to choose our factors, there is no different with the PCs that we picked before, see details on coefficient, please refer to Table 6.2 in the appendix.

6.2.2 Score

We got each factor's score of a country when we multiply one row with rotated factor pattern matrix (Figure 6.2 in appendix). Based on every factor of every country has score, we multiply each factor score with their eigenvalue to get weighted total score of every country.

We ranked all country based on their total score and table 6.2 show us the top 6 countries that have the highest score. From 1 to 6 are Luxembourg, China, Qatar, Bahrain, Iceland, United States. (Full table see table 6.3 in appendix)

6.3 Summary

We are surprised that the top one country based on our scoring will be Luxembourg, and we tried to find the reason. We focus on first five factors since their eigenvalue are relative larger than the others. We found that for Luxembourg, scores of factor1 and factor4 tend to be big, which are 1.987 and 4.642, we then referred to the coefficients of these two factors.

For factor1, top 5 variables are:

X43(+, urban population%)

X36(-, rural population%)

X9(+, employment in service%)

X21(+, individual using internet%)

X26(+, life expectancy at birth).

For factor4, top 4 variables are:

X41(+, trade (% of GDP))

X42(+, trade in services (% of GDP))

X33(+, population density)

X11(+, exports of goods and services)

Factor1 is more concentrated on people daily life and the quality of people, Factor4 is more concentrated on population and trade in services. The outcome means the reason why Luxembourg will be the top one in this analyze is that the people who live in Luxembourg has a better life than those in other countries, and have more percentage of trade in service compared to their GDP. More about factors meaning and ranking will be on report conclusion.

7.Conclusion

The criterions that determine a country is developed or not are the country has high GDP per capita, level of industrialization, human development index, and indices for life expectancy and education, etc.[1] So, we did PCA first to determine 12 PCs should be retained to explain most information of data, then after dimensionality was reduced, GDP growth rate and urban population growth rate were chose to be response values to fit the multivariate multiple PCA regression. Criterion in economic can be covered by GDP growth rate, and the urban-population growth rate account for other criterions since index of human development, life expectancy and education should be higher in urban area than rural area. We found that besides the GNI, high-tech export did a large impact on GDP growth rate and top variables to explain urban population growth rate are death rate, forest area and labor force. After clustering, we also found the patterns are related to population and economic area than divide countries into several parts. Finally, we ranked all countries according to their total score by using factor analysis which score 12 factors'. The first four factors are mainly focus on human living quality, economic environment, index of population and trade in services.

Our result indicates that if a country gets higher score, people in that country have a better living environment, most percentage of total population are living in urban area, growing economic and advanced service industry.

8.Reference

1.https://en.wikipedia.org/wiki/Developed_country

2. Rencher, Alvin C., and William F. Christensen. Methods of Multivariate Analysis. Wiley, 2012.

[1] https://en.wikipedia.org/wiki/Developed_country