

Term Project

- 사회경제적, 건강적 후원이 필요한 후진국 분류 -

제출일 : 2020년 7월 3일
통계학과 201811526 이은주

[목차]

I. Abstract	p. 2
II. Introduction	p. 2
1. 분석 동기 및 목적	p.2
2. 데이터 설명	p.2
III. Analysis and Interpretations	p. 3
1. Multivariate Normality	p.3
2. PCA	p.4
3. PCFA	p.7
4. CA	p.10
IV. Conclusion	p. 13
V. References	p. 13
VI. R Code	p. 13

I. Abstract

167개의 국가에 대한 사회경제적 요인과 건강적 요인의 특성을 파악함으로써 지원이 필요로 한 국가들을 선별하기 위해 PCA(Principal Component Analysis), FA(Factor Analysis), CA(Cluster Analysis)를 이용하여 분석한다.

II. Introduction

1. 분석 동기 및 목적

현재, 우리나라는 한강의 기적이라 불리는 경제적 성장과 의료 시설의 발전 등을 통해 과거와 달리 빈곤과 질병 등으로부터 벗어나 이제는 선진국으로써 한 발 앞서 나아가고 있다. 하지만 지구 반대편 어디에선가는 여전히 깨끗한 물 한 모금 마시지 못하고 빈곤과 싸우며 지내고 있는 많은 국가들이 있다. 이에 따라 인도적 관점에서 후진국에 대한 지원과 도움이 필요로 하지만 현실적으로 모든 국가에게 원조하는 것은 사실상 힘들다. 이에 따라 각 국가의 사회경제적 요인과 건강적 요인에 대한 다변량 데이터를 분석함으로써 각 국가들이 어떤 특성을 지니고 있는지 확인하고 이를 토대로 어떤 국가들이 가장 지원이 필요로 한지 선별하도록 한다.

2. 데이터 설명

데이터의 전체적인 형태와 각 변수에 대한 설명은 다음과 같다.

```
> summary(data)
```

	country	child_mort	exports	health	imports
Afghanistan	: 1	Min. : 2.60	Min. : 0.109	Min. : 1.810	Min. : 0.0659
Albania	: 1	1st Qu.: 8.25	1st Qu.: 23.800	1st Qu.: 4.920	1st Qu.: 30.2000
Algeria	: 1	Median : 19.30	Median : 35.000	Median : 6.320	Median : 43.3000
Angola	: 1	Mean : 38.27	Mean : 41.109	Mean : 6.816	Mean : 46.8902
Antigua and Barbuda	: 1	3rd Qu.: 62.10	3rd Qu.: 51.350	3rd Qu.: 8.600	3rd Qu.: 58.7500
Argentina	: 1	Max. : 208.00	Max. : 200.000	Max. : 17.900	Max. : 174.0000
(Other)	: 161				
	income	inflation	life_expec	total_fer	gdpp
Min. :	609	Min. : -4.210	Min. : 32.10	Min. : 1.150	Min. : 231
1st Qu.:	3355	1st Qu.: 1.810	1st Qu.: 65.30	1st Qu.: 1.795	1st Qu.: 1330
Median :	9960	Median : 5.390	Median : 73.10	Median : 2.410	Median : 4660
Mean :	17145	Mean : 7.782	Mean : 70.56	Mean : 2.948	Mean : 12964
3rd Qu.:	22800	3rd Qu.: 10.750	3rd Qu.: 76.80	3rd Qu.: 3.880	3rd Qu.: 14050
Max. :	125000	Max. : 104.000	Max. : 82.80	Max. : 7.490	Max. : 105000

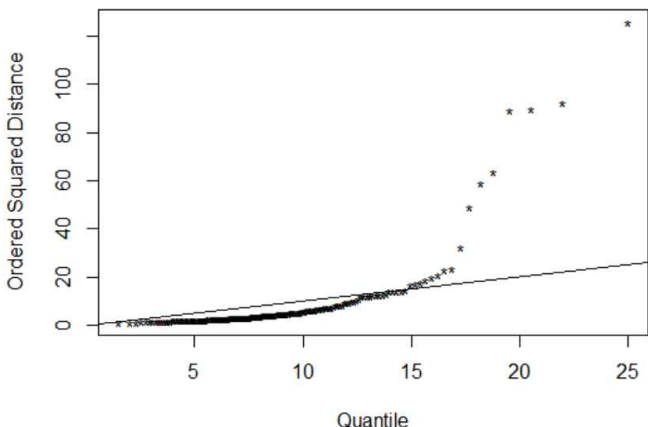
country	나라명
child_mort	1000명 당 5세 미만 아동 사망 수
exports	상품과 서비스의 수출 (총 GDP의 %age)
health	총 건강 지출 (총 GDP의 %age)
imports	상품과 서비스의 수입 (총 GDP의 %age)
income	1인당 순이익
inflation	총 GDP의 연간 성장률 측정
life_expec	현재의 사망률 패턴이 유지될 시, 새로 태어난 아이가 살 수 있는 평균 수
total_fer	현재의 연령별 임신율이 유지될 시, 각 여성에게 태어날 아이들의 수
gdpp	1인당 GDP (총 GDP를 총인구로 나눈 값)

III. Analysis and Interpretations

- * 전체 데이터에서 country 변수를 제거한 후 각 행의 이름으로 지정하여 분석하였습니다.
- * exports, health, imports 변수의 경우 백분을 값으로 국가의 인구수에 따라 영향 (ex. 수출과 수입이 높음에도 불구하고 인구수가 많아 적은 수출과 수입을 이룬다고 잘못된 판단) 을 받으므로 실제 값을 나타낼 수 있도록 gdp를 곱한 후 100으로 나누어 주었습니다.

1. Multivariate Normality

우선 데이터가 정규성을 따르는지에 대해 확인하기 위해, Chi-square Plot을 이용한 방법과 Skewness & Kurtosis에 의한 방법을 모두 확인해보고 판단하도록 한다.

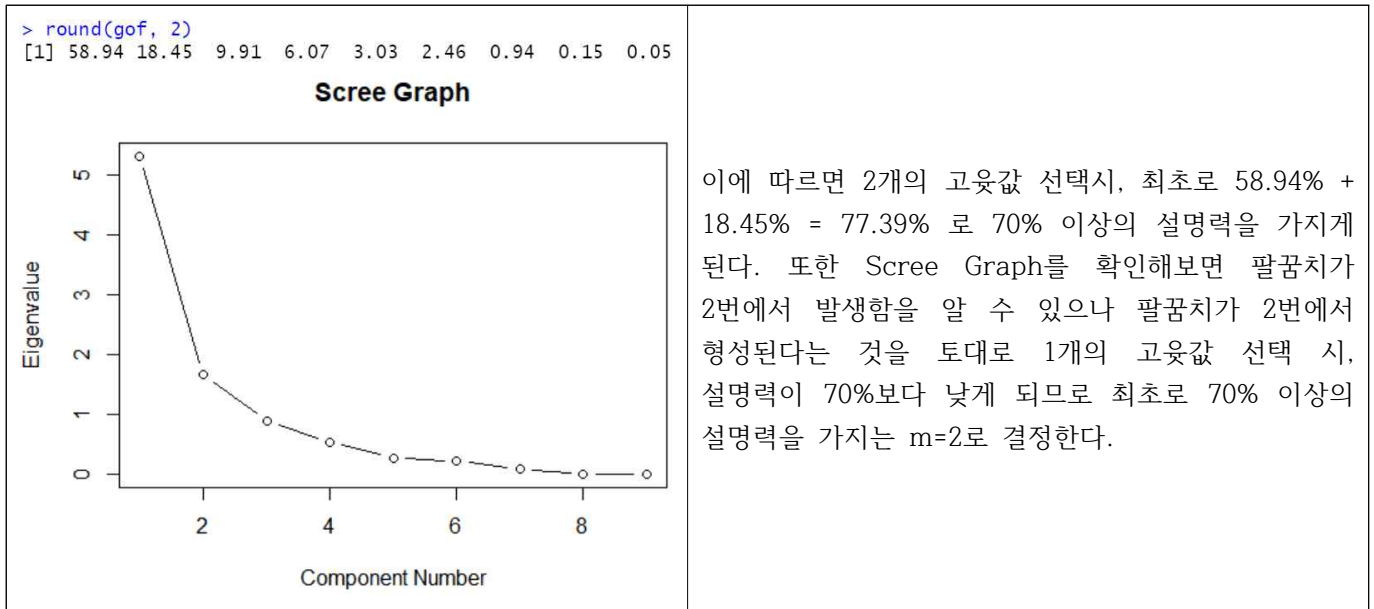
Chi-square Plot	Skewness & Kurtosis
 <pre> > rq [1] 0.759756 </pre>	<pre> > mvn(X, mvnTest="mardia", multivariatePlot = "qq") \$multivariateNormality Test Statistic p value Result 1 Mardia Skewness 6151.34732263373 0 NO 2 Mardia Kurtosis 121.073444183601 0 NO 3 MVN <NA> <NA> NO \$univariateNormality Test Variable Statistic p value Normality 1 Shapiro-wilk child_mort 0.8119 <0.001 NO 2 Shapiro-wilk exports 0.3959 <0.001 NO 3 Shapiro-wilk health 0.6069 <0.001 NO 4 Shapiro-wilk imports 0.4101 <0.001 NO 5 Shapiro-wilk income 0.7713 <0.001 NO 6 Shapiro-wilk inflation 0.6163 <0.001 NO 7 Shapiro-wilk life_expec 0.9264 <0.001 NO 8 Shapiro-wilk total_fer 0.8722 <0.001 NO 9 Shapiro-wilk gdp 0.6965 <0.001 NO </pre>
<p>Chi-square Plot을 통해 점들의 직선성을 확인함으로써 다변량 정규성에 대해 판단해 보면, 점들이 갈수록 직선으로부터 크게 벗어남을 알 수 있다. 더불어 상관계수 rq값을 확인해보면 0.759756으로 0.95 이상이 되지 않는 것을 보아 정규성을 따르지 않는다고 판단할 수 있다.</p>	<p>왜도와 첨도를 통해 다변량 정규성을 확인해 보면, 귀무가설 : “다변량 정규성을 따른다”에 대한 p-value가 0으로 귀무가설을 완전히 기각할 수 있는 것을 알 수 있다. 즉, 다변량 정규성을 따르지 않는다는 것을 알 수 있다.</p>

이에 따라 데이터는 다변량 정규성을 따르지 못한다고 할 수 있으며, 이에 따라 Factor Analysis에서 MLFA (최대우도인자분석)은 시행하지 못한다.

2. PCA

* 데이터 확인 시, 변수별로 단위가 다르므로 상관행렬 R을 이용하여 PCA를 진행하였습니다.

먼저 상관행렬 R을 구한 후, R에 대한 스펙트럼분해를 통해 고유값과 고유벡터를 구한다. 이후 고유값의 총합에서 70% 이상 설명비율의 합을 갖는 m개의 고유값을 선택하기 위해 고유값의 설명비율과 Scree Graph를 확인해보면 다음과 같다.



이에 따라 2개의 고유값에 대응하는 고유벡터를 활용하여 주성분 P1, P2를 구해보고 각각의 주성분을 해석해보면 다음과 같다.

```
> V2
```

	PC1	PC2
child_mort	0.32	-0.48
exports	-0.34	-0.40
health	-0.36	-0.16
imports	-0.34	-0.37
income	-0.38	-0.13
inflation	0.14	-0.22
life_expec	-0.34	0.37
total_fer	0.30	-0.46
gdpp	-0.40	-0.20

① PC1

우선적으로 제1주성분의 경우, 뚜렷하게 강한 양의 관계와 음의 관계를 보이는 변수가 없어 주성분 해석이 어려우나 child_mort와 total_fer과 양의 관계를 가지고, (inflation의 경우 매우 낮은 양의 상관관계를 지니므로 제외한다.) 나머지는 음의 관계를 가지는 것으로 보아 PC1이 양의 값을 가지는 경우 여성에게 태어날 아이들의 수는 많지만 태어났음에도 5세가 되기 전, 아이들이 사망한 수가 많고 수출, 수입, 건강지출, 태어난 아이가 살 수 있는 평균 수, GDP가 낮으므로 후진국이라고 할 수 있다.

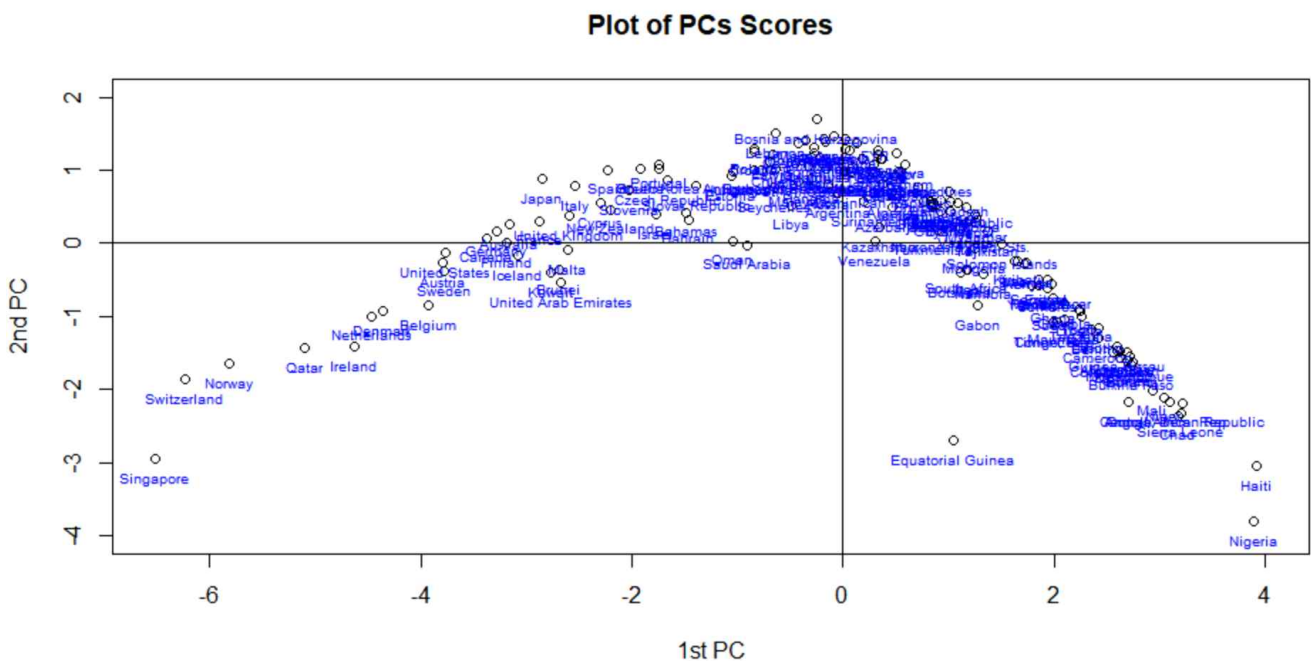
즉, PCA이 양의 값일 경우 후진국, 음의 값일 경우 선진국이라고 할 수 있다.

② PC2

제 2주성분의 경우에도 뚜렷하진 않지만, PC2와 변수들의 양, 음의 관계를 보고 주성분을 해석해보면, PC2가 양의 값을 가질 때, 태어날 아이들 수와 5세 미만의 아동 사망 수가 적고, 태어난 아이가 살 수 있는 평균 수가 높음을 알 수 있다. 또한, 수출과 수입이 적은 것을 알 수 있는데, 이는 경제적인 무역활동이 발전되지 않았으나 건강적 요인이 좋은 것을 보아 수출, 수입 의존도가 낮고 국가 내에서 자생하는 경우로 예상할 수 있다. (이때, 태어날 아이들 수가 적으므로 아동 사망 수가 적은 것이라고 생각할 수 있으나 child_mort는 1000명당 아동 사망 수로 동일한 단위에서 비롯되므로 인과관계로 고려하지 않는다.)

또한, 주성분점수를 얻어 새로운 다변량 자료로 보고 주성분 해석에 따라 각 국가의 특성을 확인한다.

1st PC and 2nd PC Plot of PCs Scores

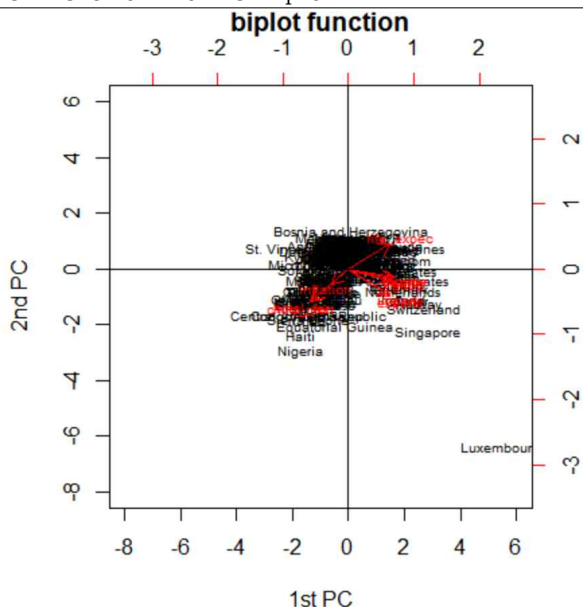


위의 주성분 해석을 바탕으로 위의 주성분점수 그림을 확인해보면, 제1주성분 중심으로 Nigeria, Haiti, Central African Republic 등이 양의 값을 가지므로 각 여성에게 태어날 아이들의 수는 많으나 태어난 이후 5세 미만 아이들이 사망한 수가 많고 경제적으로도 부족한 후진국임을 알 수 있다. 반면에, Singapore, Switzerland, Norway 등의 경우 태어난 아이가 살 수 있는 평균 수가 높고 경제적으로도 우수한 선진국임을 알 수 있다.

더불어 제2주성분 중심으로 주성분점수 그림을 확인해보면, Bosnia and Herzegovina, Lebanon 등이 양의 값을 가지므로 수출, 수입에 대한 의존도가 낮고 Singapore, Nigeria, Haiti 등은 수출, 수입에 대한 의존도가 높음을 알 수 있다. 즉, Singapore, Nigeria, Haiti 등과 같은 경우 주 경제적 활동이 무역활동임을 알 수 있다. 또한, 이 그림을 통해 제2주성분을 제1주성분과 비교하여 고려해보았을 때, 경제적 요인이 건강적 요인과 독립적이라는 것에 대해 고려해볼 수 있다. (ex. Haiti, Nigeria (후진국)의 주요 경제적 활동이 무역 활동임에도 불구하고 아이의 사망 수가 높다.)

더불어 주성분점수에 대한 그림에 대한 해석을 각 변수들 관점에서 확인하기 위해 Biplot을 보면 다음과 같다.

1st PC and 2nd PC Biplot

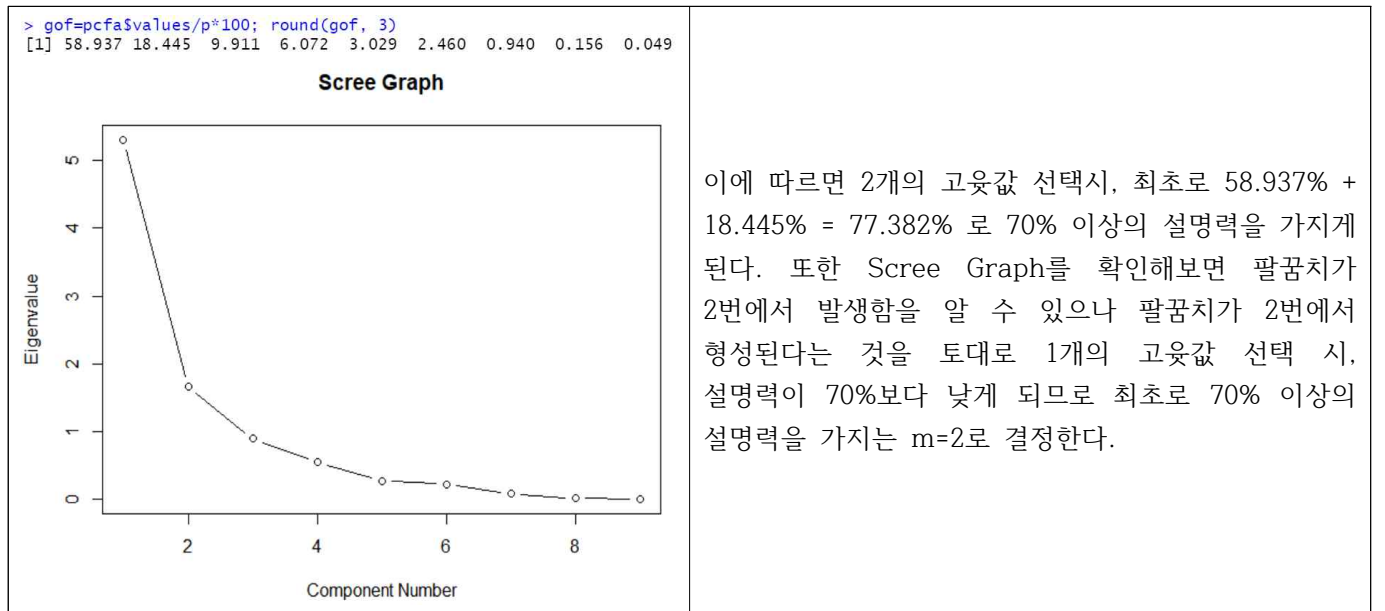


*이때, 각 개체가 Luxembourg를 제외하고 각 국가들이 모여있으므로 이에 맞게 xlim과 ylim을 조절하여 다시 확인 하였습니다.

3. PCFA

- * Factor Analysis에는 PCFA와 MLFA가 있으나 MLFA는 데이터가 정규성을 만족해야 하므로 생략합니다.
- * PCA와 같은 이유로 Z를 이용하였고 공통인자의 해석을 더 명확하게 하기 위해 인자회전에 따른 PCFA를 실시하였습니다.

PCA와 마찬가지로 먼저 인자의 개수를 정하기 위하여 고윳값의 설명비율과 Scree Graph를 확인해보면 다음과 같다.



이에 따라 2개의 인자 f1, f2를 구하고 해석해보면 다음과 같다.

```
> L=pcfa$loading[,1:2];
               RC1    RC2
child_mort -0.207  0.930
exports    0.938 -0.072
health     0.778 -0.342
imports    0.921 -0.102
income     0.796 -0.400
inflation  -0.089  0.427
life_expec 0.340 -0.859
total_fer  -0.195  0.894
gdpp       0.889 -0.354
```

① f1 - RC2
첫 번째 인자의 경우 exports, health, imports, imcome, gdpp와 강한 양의 관계에 있는 것을 알 수 있다. 즉, 이는 경제적 요인과 관련된 인자로 f1이 양의 값을 가지는 경우 높은 경제적 수준을 가지고 있는 국가라고 판단할 수 있다. 또한, 음의 값을 가지는 경우에는 경제적 수준이 낮은 국가라고 할 수 있다.
② f2 - RC4
두 번째 인자의 경우 child_mort, total_fer과 강한 양의 관계에 있으며 또한 life_expec과 강한 음의 관계를 가진다. 이에 따라 f2가 건강적 요인과 관련된 인자로 양의 값을 가지는 경우 태어나는 아이가 많으나 5세가 되기 전 사망하는 수가 많으며 음의 값을 가지는 경우 태어난 아이가 살 수 있는 평균 수가 높다고 할 수 있다. 즉, f2가 양의 값을 가지는 경우 의료적 수준이 매우 낮으며 음의 값을 가지는 경우 의료적 수준이 높다고 할 수 있다.

-> PCA와 비교하였을 때 인자에 대한 해석이 매우 수월해진 것을 알 수 있다.

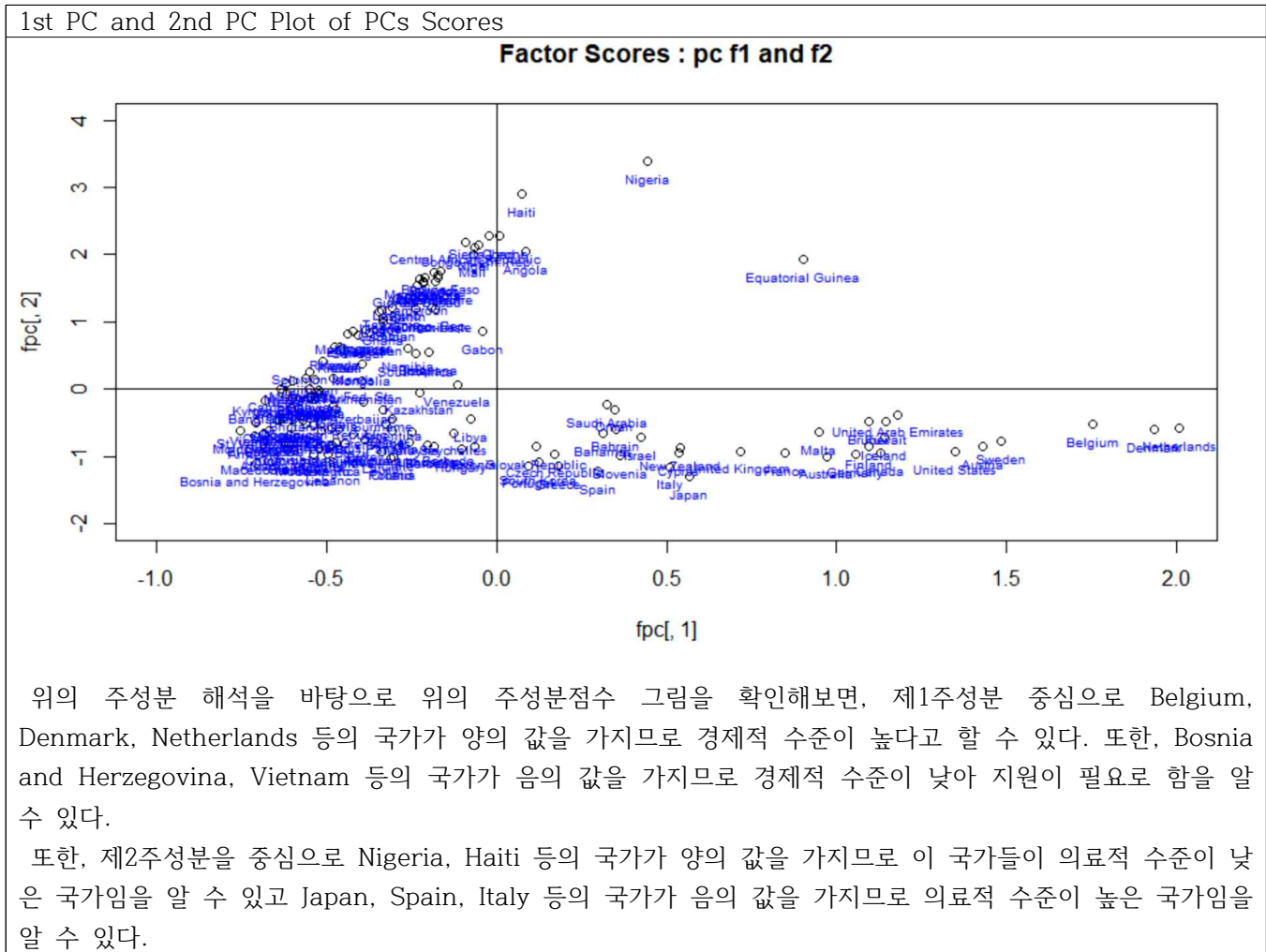
이때, 인자적재그림과 인자점수그림을 얻기 전, 2개의 인자를 선택한 모델이 적절한지 확인하기 위하여 공통성, 특정분산, 잔차행렬을 구해본다.

1) 공통성 - 1에 가까울수록 선택한 요인이 원데이터의 변수들을 잘 설명하고 있다고 할 수 있다.									
<pre>> round(diag(L%*%t(L)), 3)</pre>									
child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	
0.908	0.886	0.722	0.859	0.793	0.190	0.854	0.837	0.915	
inflation 변수에 대한 설명력이 낮다고 볼 수 있으나 대체로 1에 가까우므로 모델이 적절하다고 할 수 있다.									
2) 특정분산 - 0에 가까울수록 선택한 요인이 원데이터의 변수들을 잘 설명하고 있다고 할 수 있다.									
<pre>> round(Psi,2)</pre>									
child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	
0.09	0.11	0.28	0.14	0.21	0.81	0.15	0.16	0.08	
특정분산 확인시에도 inflation 변수가 1에 가까우나 이를 제외하고 모두 0에 매우 가까우므로 2개의 인자를 선택한 모델이 적절하다고 할 수 있다.									
3) 잔차행렬 - 대각원소가 0이고 비대각원소가 전반적으로 매우 작으면 인자모형이 적절하다고 할 수 있다.									
<pre>> round(Rm, 2)</pre>									
	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
child_mort	0.00	-0.04	0.05	-0.03	0.01	-0.13	-0.02	-0.02	0.03
exports	-0.04	0.00	-0.14	0.12	-0.05	-0.03	0.00	-0.04	-0.09
health	0.05	-0.14	0.00	-0.11	-0.06	-0.04	-0.01	0.05	0.10
imports	-0.03	0.12	-0.11	0.00	-0.10	-0.05	0.00	-0.05	-0.10
income	0.01	-0.05	-0.06	-0.10	0.00	0.09	0.00	0.01	0.05
inflation	-0.13	-0.03	-0.04	-0.05	0.09	0.00	0.16	-0.08	0.01
life_expec	-0.02	0.00	-0.01	0.00	0.00	0.16	0.00	0.07	-0.01
total_fer	-0.02	-0.04	0.05	-0.05	0.01	-0.08	0.07	0.00	0.03
gdpp	0.03	-0.09	0.10	-0.10	0.05	0.01	-0.01	0.03	0.00
잔차행렬 확인 시, 대각 원소가 0이며 비대각 원소가 매우 작은 것을 알 수 있으므로 인자모델이 적절하다고 판단할 수 있다.									

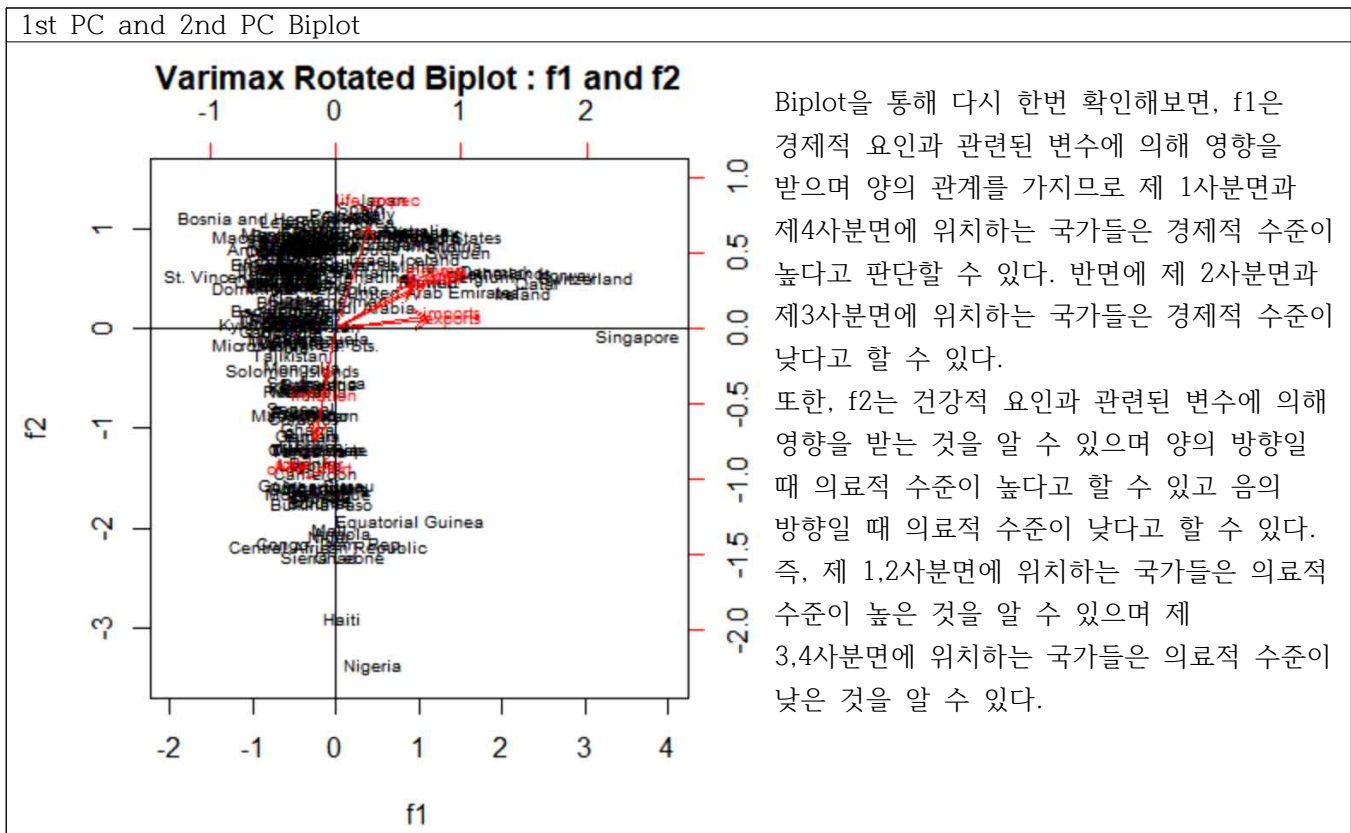
이어서 인자적재그림을 통해 2개의 인자와 원데이터의 변수의 관계를 확인해보면 다음과 같다.

1st PC and 2nd PC Plot of PC Factor Loadings	
<p style="text-align: center;">PC Factor Loadings : f1 and f2</p>	
<p>먼저 f1을 중심으로 인자적재그림을 보았을 때, 위의 인자 해석을 토대로 f1이 양의 값을 가지는 경우 경제적요인(income, gdpp, health...)이 높은 값을 가지는 것을 알 수 있다. 즉, f1이 양의 값을 가질 때 경제적 수준이 높다고 할 수 있고 음의 값을 가질 때 경제적 수준이 낮다고 할 수 있다. 이때, 각 변수에 대한 정사영 길이가 길 경우 인자에 대한 영향이 큰 것을 의미하는 데 이에 따라 f1은 inflation과 건강적 요인에 따른 영향이 거의 없다고 볼 수 있다.</p> <p>또한, f2를 중심으로 인자적재그림을 보았을 때, 마찬가지로 f2 인자는 경제적 요인에 대한 영향이 거의 없음을 알 수 있으며 f2가 양의 값을 가지는 경우 태어나는 아이는 많으나 5세가 되기 전 사망하는 수가 많고 음의 값을 가지는 경우 태어난 아이가 살 수 있는 평균 수가 높다고 할 수 있다. 즉, f2가 양의 방향일 때 의료적 수준이 낮으며 음의 방향일 때 의료적 수준이 높다고 할 수 있다.</p> <p>이에 따라 인자적재그림 확인 시 인자적재값에 따른 해석을 따를 수 있다.</p>	

또한, 인자점수그림을 통해 인자의 해석에 따라 각 국가의 특성을 확인한다.



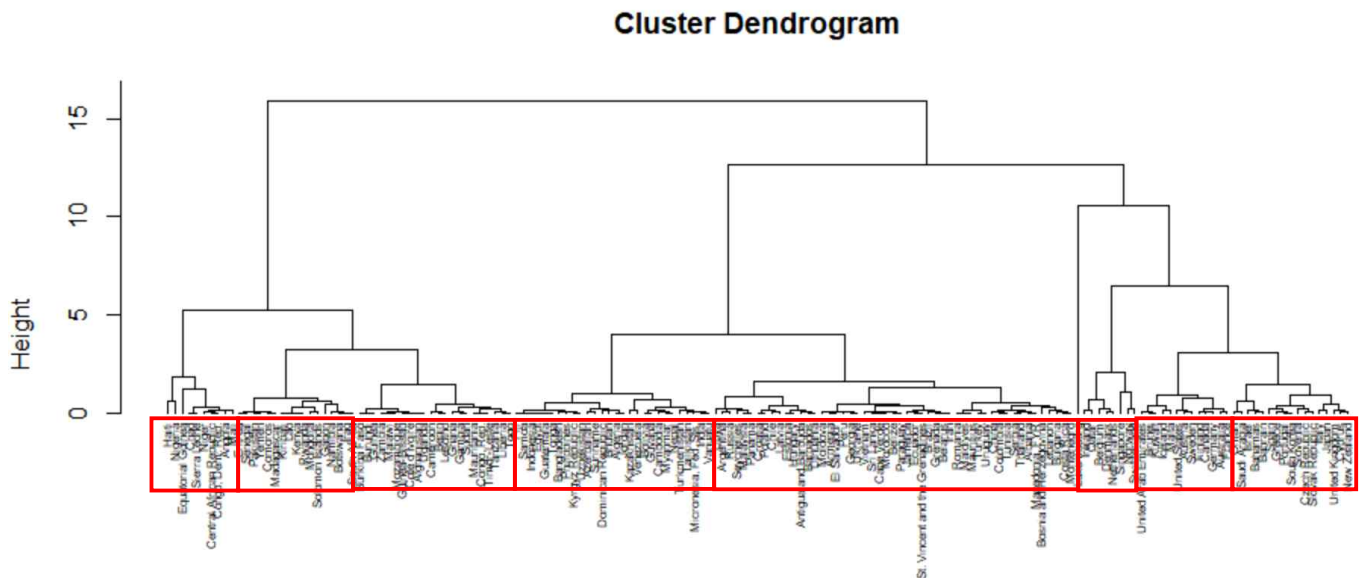
더불어 인자행렬도를 통해 각 국가와 변수들의 관계를 확인해보면 다음과 같다.



4. CA

PCA와 PCFA의 결과에 따라 경제적 요인과 건강적 요인을 구분하였고 이에 따라 경제적 수준이 낮은 국가와 의료적 수준이 낮은 국가를 분류하여 인도적 지원을 하고자 한다. 이에 따라 군집분석을 실시하며 이때, PCA보다 PCFA에 따른 2개의 인자 모델이 뚜렷한 해석이 수월했으므로 앞에서 PCFA를 통해 구한 인자 점수를 새로운 다변량 자료로 보고 이를 이용하여 군집분석을 실시한다. 군집분석에는 계층 군집분석과 비계층 군집분석이 있으며 이에 따라 계층군집분석의 대표적인 방법 중 와드연결법과 비계층 군집분석 중 K-평균법을 실시한다.

① Hierarchical clustering methods - Ward linkage



약 8개의 군집으로 나타나는 것을 알 수 있으나 개체의 수가 많아 군집별 특성을 파악하기 어렵다. 이에 따라 K-평균법에 따른 비계층 군집분석에 의하여 군집을 나누고 이에 대한 해석을 하도록 한다.

② Non-hierarchical clustering methods - K_means method

```

* Among all indices:
* 4 proposed 2 as the best number of clusters
* 5 proposed 3 as the best number of clusters
* 8 proposed 4 as the best number of clusters
* 1 proposed 5 as the best number of clusters
* 4 proposed 8 as the best number of clusters
* 1 proposed 10 as the best number of clusters

```

***** Conclusion *****

* According to the majority rule, the best number of clusters is 4

먼저, 군집의 수를 결정하기 위하여 모든 평가지표를 이용할 시, 4개의 군집수를 선택하는 것이 가장 적절함을 알 수 있다. 이에 따라 군집수를 4개로 결정하고 K-평균법을 이용하여 군집을 나눈다면 다음과 같다.

	rownames.fpc.	cluster
Australia	Australia	1
Austria	Austria	1
Belgium	Belgium	1
Brunei	Brunei	1
Canada	Canada	1
Denmark	Denmark	1
Finland	Finland	1
Germany	Germany	1
Iceland	Iceland	1
Ireland	Ireland	1
Kuwait	Kuwait	1
Luxembourg	Luxembourg	1
Malta	Malta	1
Netherlands	Netherlands	1
Norway	Norway	1
Qatar	Qatar	1
Singapore	Singapore	1
Sweden	Sweden	1
Switzerland	Switzerland	1
United Arab Emirates	United Arab Emirates	1
United States	United States	1

	rownames.fpc.	cluster
Afghanistan	Afghanistan	2
Angola	Angola	2
Benin	Benin	2
Burkina Faso	Burkina Faso	2
Burundi	Burundi	2
Cameroon	Cameroon	2
Central African Republic	Central African Republic	2
Chad	Chad	2
Congo, Dem. Rep.	Congo, Dem. Rep.	2
Congo, Rep.	Congo, Rep.	2
Cote d'Ivoire	Cote d'Ivoire	2
Equatorial Guinea	Equatorial Guinea	2
Gambia	Gambia	2
Guinea	Guinea	2
Guinea-Bissau	Guinea-Bissau	2
Haiti	Haiti	2
Lesotho	Lesotho	2
Liberia	Liberia	2
Malawi	Malawi	2
Mali	Mali	2
Mauritania	Mauritania	2
Mozambique	Mozambique	2
Niger	Niger	2
Nigeria	Nigeria	2
Sierra Leone	Sierra Leone	2
Sudan	Sudan	2
Tanzania	Tanzania	2
Timor-Leste	Timor-Leste	2
Togo	Togo	2
Uganda	Uganda	2
Zambia	Zambia	2

	rownames.fpc.	cluster
Azerbaijan	Azerbaijan	3
Bangladesh	Bangladesh	3
Bolivia	Bolivia	3
Botswana	Botswana	3
Cambodia	Cambodia	3
Comoros	Comoros	3
Egypt	Egypt	3
Eritrea	Eritrea	3
Fiji	Fiji	3
Gabon	Gabon	3
Ghana	Ghana	3
Guatemala	Guatemala	3
Guyana	Guyana	3
India	India	3
Indonesia	Indonesia	3
Iraq	Iraq	3
Kazakhstan	Kazakhstan	3
Kenya	Kenya	3
Kiribati	Kiribati	3
Kyrgyz Republic	Kyrgyz Republic	3
Lao	Lao	3
Madagascar	Madagascar	3
Micronesia, Fed. Sts.	Micronesia, Fed. Sts.	3
Mongolia	Mongolia	3
Myanmar	Myanmar	3
Namibia	Namibia	3
Nepal	Nepal	3
Pakistan	Pakistan	3
Philippines	Philippines	3
Rwanda	Rwanda	3
Samoa	Samoa	3
Senegal	Senegal	3
Solomon Islands	Solomon Islands	3
South Africa	South Africa	3
Tajikistan	Tajikistan	3
Tonga	Tonga	3
Turkmenistan	Turkmenistan	3
Uzbekistan	Uzbekistan	3
Vanuatu	Vanuatu	3
Venezuela	Venezuela	3
Yemen	Yemen	3

	rownames.fpc.	cluster
Albania	Albania	4
Algeria	Algeria	4
Antigua and Barbuda	Antigua and Barbuda	4
Argentina	Argentina	4
Armenia	Armenia	4
Bahamas	Bahamas	4
Bahrain	Bahrain	4
Barbados	Barbados	4
Belarus	Belarus	4
Belize	Belize	4
Bhutan	Bhutan	4
Bosnia and Herzegovina	Bosnia and Herzegovina	4
Brazil	Brazil	4
Bulgaria	Bulgaria	4
Cape Verde	Cape Verde	4
Chile	Chile	4
China	China	4
Colombia	Colombia	4
Costa Rica	Costa Rica	4
Croatia	Croatia	4
Cyprus	Cyprus	4
Czech Republic	Czech Republic	4
Dominican Republic	Dominican Republic	4
Ecuador	Ecuador	4
El Salvador	El Salvador	4
Estonia	Estonia	4
France	France	4
Georgia	Georgia	4
Greece	Greece	4
Grenada	Grenada	4
Hungary	Hungary	4
Iran	Iran	4
Israel	Israel	4
Italy	Italy	4
Jamaica	Jamaica	4
Japan	Japan	4
Jordan	Jordan	4
Latvia	Latvia	4
Lebanon	Lebanon	4
Libya	Libya	4
Lithuania	Lithuania	4
Macedonia, FYR	Macedonia, FYR	4
Malaysia	Malaysia	4

	rownames.fpc.	cluster
Maldives	Maldives	4
Mauritius	Mauritius	4
Moldova	Moldova	4
Montenegro	Montenegro	4
Morocco	Morocco	4
New Zealand	New Zealand	4
Oman	Oman	4
Panama	Panama	4
Paraguay	Paraguay	4
Peru	Peru	4
Poland	Poland	4
Portugal	Portugal	4
Romania	Romania	4
Russia	Russia	4
Saudi Arabia	Saudi Arabia	4
Serbia	Serbia	4
Seychelles	Seychelles	4
Slovak Republic	Slovak Republic	4
Slovenia	Slovenia	4
South Korea	South Korea	4
Spain	Spain	4
Sri Lanka	Sri Lanka	4
St. Vincent and the Grenadines	St. Vincent and the Grenadines	4
Suriname	Suriname	4
Thailand	Thailand	4
Tunisia	Tunisia	4
Turkey	Turkey	4
Ukraine	Ukraine	4
United Kingdom	United Kingdom	4
Uruguay	Uruguay	4
Vietnam	Vietnam	4

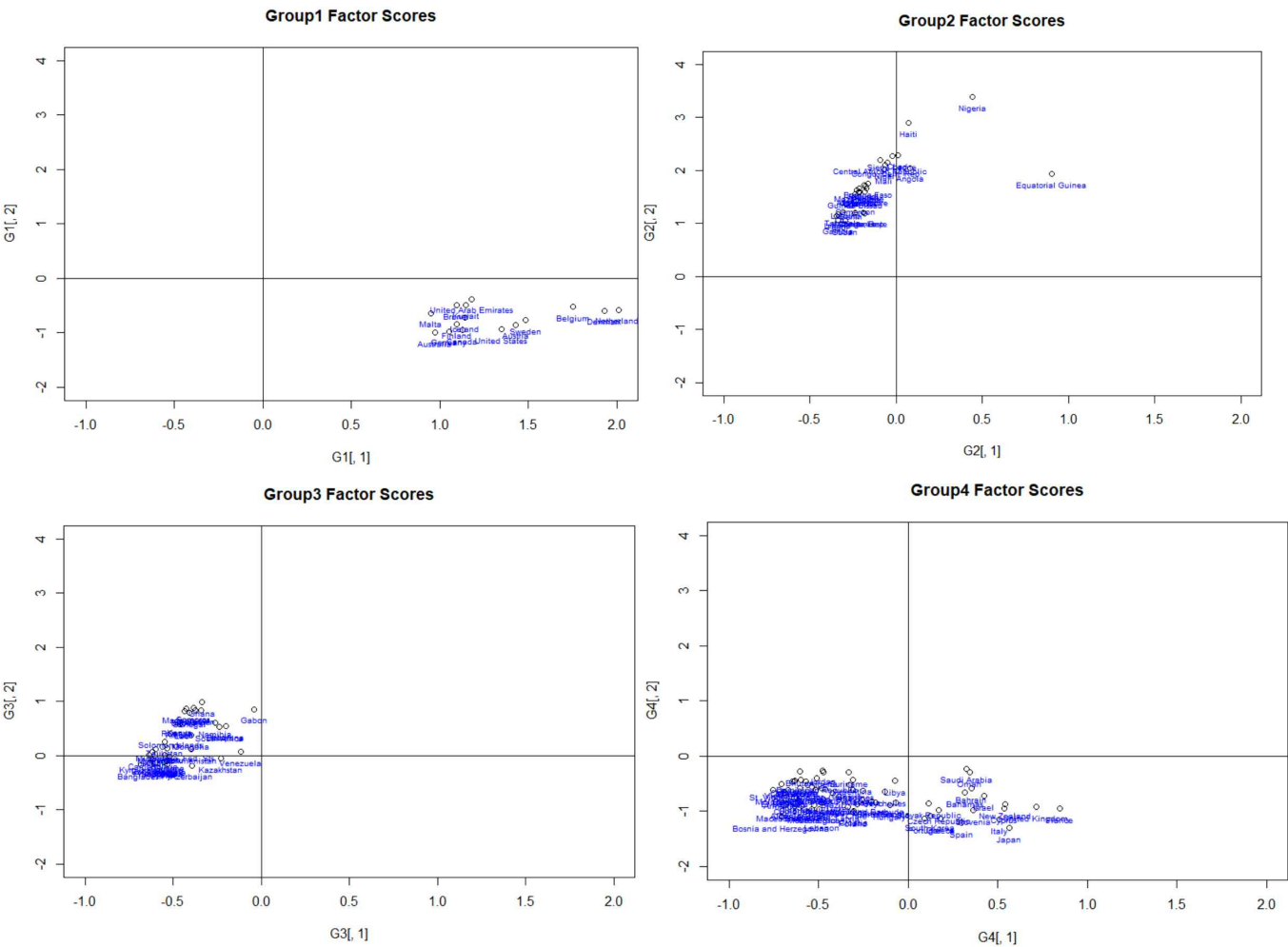
각 군집의 특성을 파악하기 위해 각 군집별 f1과 f2의 평균을 확인하고 [3. PCFA]에서의 f1과 f2 인자의 해석을 바탕으로 각 군집의 특성을 알아보면 다음과 같다.

```
> aggregate(fpc, by=list(kmeans$cluster), FUN=mean)
```

	Group.1	RC1	RC2
1	1	2.0051176	-0.5115251
2	2	-0.1225044	1.7139661
3	3	-0.4492119	0.2958234
4	4	-0.2688128	-0.7367525

Group1	G1은 다른 군집에 비해 f1값이 매우 높고 f2값이 낮은 것을 알 수 있다. 이에 따라 f1값이 경제적 수준이 높으며 의료적 수준이 높은 군집이라고 할 수 있다. 즉, 경제적, 의료적 수준이 모두 발전된 국가들의 군집임을 알 수 있다.
Group2	G2는 f1값이 G3와 G4에 비해 높고 f2값이 매우 높은 것을 보아 경제적 수준은 상대적으로 적절하나 의료적 수준이 매우 낮은 국가들이라고 할 수 있다. 즉, 의료적 지원이 필요로한 국가들이다.
Group3	G3는 f1값이 매우 낮고 f2값이 약간 높으므로 의료적 수준은 상대적으로 적절하나 경제적 수준이 매우 낮은 국가들이라고 할 수 있으므로 경제적 차원에서 도움이 필요로한 군집이다.
Group4	G4는 f1값이 낮고 특히 f2값이 매우 낮은 것을 보아 경제적 수준은 낮으나 의료적 수준이 매우 높은 국가들이라고 할 수 있다.

추가적으로 그림을 통해 각 군집에 대해 알아보면 다음과 같다.



군집별 Factor Scores를 통해 군집의 특성이 각 군집별 평균으로부터 추론한 특성과 동일함을 알 수 있다.

IV. Conclusion

다차원 데이터에서 각 개체들, 즉 국가들의 특성을 각 변수마다 파악하고 구별하기는 어렵다. 이에 따라 PCA(Principal Component Analysis)와 FA(Factor Analysis)를 통해 차원 축소를 이루게 된다면 데이터에 대해 더욱 수월하게 이해할 수 있게 된다. 앞서 진행된 PCA와 PCFA를 통해서 9개의 차원으로부터 2차원으로 차원 축소하여 각 변수들을 경제적 요인과 건강적 요인으로 구별할 수 있었다. 특히 인자회전을 통해 시행한 각 인자에 대한 해석이 더욱 명확해져 이를 토대로 각 국가의 특성을 파악할 수 있게 되었다.

더 나아가 데이터를 국가의 특성이 유사한 군집으로 나누어 실질적으로 인도적 지원이 필요로 한 군집을 찾도록 한다. 이에 따라 CA 결과를 확인해 보면, Group2 군집 내의 국가들에게는 의료적 지원 시스템을 구축하도록 하고 Group3 군집 내의 국가들은 앞으로 자유 무역과 국제화 등을 통해 글로벌한 경제적 시스템을 구축할 수 있는 방안을 마련하도록 한다.

즉, 이처럼 다변량 분석에서 각 국가의 특성을 차원축소를 통해 쉽게 구분하고 그 특성에 따른 군집을 선별할 수 있게 되어 실제로 꼭 필요한 국가에게 맞는 도움을 줄 수 있게 된다.

V. References

데이터 출처 : <https://www.kaggle.com/hellbuoy/pca-kmeans-hierarchical-clustering>

VI. R Code

```
####Multivariate Statistics (I) Term Project
####2018111526 이은주
setwd("C:\\Users\\eunju\\Desktop\\junior\\TT10.30 다변량통계학\\Term Project")

####data
data = read.csv("Underdeveloped Countries.csv", header=TRUE)
head(data)
summary(data)

X = data[,-1]
rownames(X) = data[,1]

X$exports = X$exports*X$gdpp/100
X$imports = X$imports*X$gdpp/100
X$health = X$health*X$gdpp/100

n=dim(X)[1]; p=dim(X)[2]
Z=scale(X, scale=T)
```

###1. Multivariate Normality

```
S=cov(X)
xbar=colMeans(X)
m=mahalanobis(X, xbar, S)
m=sort(m)
id=seq(1, n)
pt=(id-0.5)/n
q=qchisq(pt, p)
plot(q, m, pch="*", xlab="Quantile", ylab="Ordered Squared Distance")
abline(0, 1)
rq=cor(cbind(q, m))[1,2]
rq
```

```
library(MVN)
mvn(X, mvnTest="mardia", multivariatePlot = "qq")
```

###2. PCA

```
R=round(cor(X),3)
R
eigen=eigen(R)
round(eigen$values,2)
V=round(eigen$vectors,2)
V

gof=eigen$values/sum(eigen$values)*100
round(gof, 2)
plot(eigen$values, type="b", main="Scree Graph",
      xlab="Component Number", ylab="Eigenvalue")
V2=V[,1:2]
rownames(V2) = colnames(X); colnames(V2) = c("PC1","PC2")
V2
```

#PCs scores

```
PC=Z%*%V2
round(PC, 3)

plot(PC[,1], PC[,2], main="Plot of PCs Scores", xlab="1st PC", ylab="2nd PC",
      xlim=c(-6.5,4), ylim=c(-4,2))
text(PC[,1], PC[,2], labels=rownames(X), cex=0.6, col="blue", pos=1)
abline(v=0, h=0)
```

#Biplot

```
n= nrow(X)
joinnames= c(rownames(X),colnames(X))
Z=scale(X, scale=T); svd.Z <- svd(Z)
U <- svd.Z$u
```



```

V <- svd.Z$v
D <- diag(svd.Z$d)
G <- (sqrt(n-1)*U)[,1:2]
H <- (sqrt(1/(n-1))*V*%D)[,1:2]
C <- rbind(G, H)
rownames(G)<-rownames(X)
rownames(H)<-colnames(X)
rownames(C)<-joinnames

eig <- (svd.Z$d)^2
per <- eig/sum(eig)*100; per
gof <- sum(per[1:2])

lim<-range(pretty(G))
biplot(G[,1:2],H[,1:2], xlab="1st PC", ylab="2nd PC", main="biplot function",
       xlim=lim,ylim=lim,cex=0.6,pch=16)
abline(v=0,h=0)
biplot(G[,1:2],H[,1:2], xlab="1st PC", ylab="2nd PC", main="biplot function",
       xlim=c(-4,4),ylim=c(-2.5,2),cex=0.6,pch=16)
abline(v=0,h=0)

###PCFA
library(psych)
pcfa<-principal(Z, nfactors=2, rotate="varimax")

round(pcfa$values, 2)
gof=pcfa$values/p*100; round(gof, 3)
plot(pcfa$values, type="b", main="Scree Graph",
     xlab="Component Number", ylab="Eigenvalue")

L=pcfa$loadings[,1:2]; round(L, 3)
round(diag(L%*%t(L)), 3)
Psi=pcfa$uniquenesses
round(Psi,2)
R=cor(X)
Rm = R-(L%*%t(L) + diag(Psi))
round(Rm, 2)

#factor loadings
lim<-range(pretty(L))
plot(L[,1], L[,2],main="PC Factor Loadings : f1 and f2",
     xlab="f1", ylab="f2", xlim=lim, ylim=lim)
text(L[,1], L[, 2], labels=rownames(L), cex=0.6, col="blue", pos=1)
abline(v=0, h=0)
arrows(0,0, L[,1], L[, 2], col=2, code=2, length=0.1)

```

```

#factor scores
fpc=pcfa$scores
round(fpc, 3)
plot(fpc[,1], fpc[,2],main="Factor Scores : pc f1 and f2",
      xlim=c(-1,2), ylim=c(-2,4))
text(fpc[,1], fpc[,2], labels=rownames(fpc), cex=0.6, col="blue", pos=1)
abline(v=0, h=0)

#Biplot
svd.Z=svd(Z)
U=svd.Z$u
V=svd.Z$v
D <- diag(svd.Z$d)
F <- (sqrt(n-1)*U)[,1:2]
L <- (sqrt(1/(n-1))*V*%D)[,1:2]
C <- rbind(F, L)
rownames(F)<-rownames(X);
rownames(L)<-colnames(X)

eig <- (svd.Z$d)^2
per <- eig/sum(eig)*100; per
gof <- sum(per[1:2])

varimax<-varimax(L)
Lt = varimax$loadings
T=varimax$rotmat; T
Ft= F*%T

biplot(Ft[,c(1,2)],Lt[,c(1,2)], xlab="f1",ylab="f2",
      main="Varimax Rotated Biplot : f1 and f2",
      xlim=c(-2,4),ylim=c(-3.5,1.5),cex=0.6,pch=16)
abline(v=0,h=0)

###CA
#Hierarchical clustering methods - Ward linkage
ds = dist(fpc, method="euclidean")
ward = hclust(ds, method="ward.D2")
plot(ward, hang=-1, labels=rownames(X), cex=0.5)

#Non-hierarchical clustering methods - K-means
library(NbClust)
all<-NbClust(fpc, distance="euclidean", min.nc = 2, max.nc = 10,
             method = "kmeans", index = "all")

kmeans <- kmeans(fpc, 4)
cluster=data.frame(rownames(fpc), cluster=kmeans$cluster)

```

```

C1=cluster[(cluster[,2]==1),]
C2=cluster[(cluster[,2]==2),]
C3=cluster[(cluster[,2]==3),]
C4=cluster[(cluster[,2]==4),]
C1:C2:C3:C4

aggregate(fpc, by=list(kmeans$cluster), FUN=mean)

###Conclusion
G1 = matrix(0,0,2)
for (i in 1:n){
  for (j in 1:dim(C1)[1]){
    if (rownames(fpc)[i]==rownames(C1)[j]){
      G1 = rbind(G1,fpc[i,])
    }
  }
}
rownames(G1) = rownames(C1)
plot(G1[,1], G1[,2],main="Group1 Factor Scores",
     xlim=c(-1,2), ylim=c(-2,4))
text(G1[,1], G1[,2], labels=rownames(G1), cex=0.6, col="blue", pos=1)
abline(v=0, h=0)

G2 = matrix(0,0,2)
for (i in 1:n){
  for (j in 1:dim(C2)[1]){
    if (rownames(fpc)[i]==rownames(C2)[j]){
      G2 = rbind(G2,fpc[i,])
    }
  }
}
rownames(G2) = rownames(C2)
plot(G2[,1], G2[,2],main="Group2 Factor Scores",
     xlim=c(-1,2), ylim=c(-2,4))
text(G2[,1], G2[,2], labels=rownames(G2), cex=0.6, col="blue", pos=1)
abline(v=0, h=0)

G3 = matrix(0,0,2)
for (i in 1:n){
  for (j in 1:dim(C3)[1]){
    if (rownames(fpc)[i]==rownames(C3)[j]){
      G3 = rbind(G3,fpc[i,])
    }
  }
}
rownames(G3) = rownames(C3)

```

```
plot(G3[,1], G3[,2],main="Group3 Factor Scores",
     xlim=c(-1,2), ylim=c(-2,4))
text(G3[,1], G3[,2], labels=rownames(G3), cex=0.6, col="blue", pos=1)
abline(v=0, h=0)

G4 = matrix(0,0,2)
for (i in 1:n){
  for (j in 1:dim(C4)[1]){
    if (rownames(fpc)[i]==rownames(C4)[j]){
      G4 = rbind(G4,fpc[i,])
    }
  }
}
rownames(G4) = rownames(C4)
plot(G4[,1], G4[,2],main="Group4 Factor Scores",
     xlim=c(-1,2), ylim=c(-2,4))
text(G4[,1], G4[,2], labels=rownames(G4), cex=0.6, col="blue", pos=1)
abline(v=0, h=0)
```