**다변량통계학 학기말과제 기말보고서**

<World Happiness Report를 활용한 국가 군집별 특징분석>

담당교수 : 김규성 교수님

제출일자 : 2022.12.19

**2019580027 정우진**

목차

**1. 서론3**

1.1. 연구 목적3

1.2. 문헌 연구 3

1.3. 데이터 설명 4

1.4 분석 방법5

1.5 결과 활용 및 기대 효과 5

**2. 본론3**

2.1. 분석 방법 소개3

2.2. 데이터 분석 및 결과 설명 3

2.3. 분석의 타당성 설명 4

**3. 결론3**

3.1. 분석 결과 요약3

3.2. 분석의 장점 및 한계점 설명3

3.3. 추가 연구사항 제안 4

참고문헌**7**

1. **서론**

**1.1 연구 목적**

매년, WRIO에서는 각 나라별로 조사한 몇 개의 변수를 바탕으로 World Happiness Report(이하 WHR)를 작성하여 배포하고 있다. 이 보고서에는 행복도에 영향을 미치는 것으로 알려진 변수들에 대한 추정치와, 이로부터 산출된 행복도 점수를 바탕으로 세계 각국의 행복도 순위를 나열한다. 본 연구에서는, 2021년 WHR을 바탕으로, 행복도, 대륙 등의 변수에 따른 행복도 또는 변수값의 특징에 대해 다변량 통계학적 기법을 사용하여 분석하고자 한다.

본 연구에서 알아보고자 하는 바는 다음과 같다.

* 행복도가 높은 나라들과 낮은 나라 간의 차이가 무엇인가? 이 때 가장 영향력이 큰 변수들에는 무엇이 있는가?
* 지정학적 요인은 각 나라의 사회, 문화를 포함한 여러가지 상황에 영향을 미친다. 각 국가가 위치한 지역별로 나타나는 차이점이 있는가? 있다면 어떤 차이가 있는가?
* 주성분 분석을 활용하여 행복도에 큰 영향을 미치는 변수 탐색

**1.2 문헌 연구**

WHR은 그 데이터가 Kaggle 등의 사이트에서 제공되고 있기 때문에, 이를 가지고 데이터 분석을 한 자료들은 다수 있으나, 이를 분석한 논문이나 연구는 알려진 것이 없었다.

**1.3 데이터 설명**

연구에서 이용한 데이터는 2021년 WHR 데이터로, WRIO에서 발표한 내용 데이터를 Kaggle 사이트에서 구했다.

연구에 사용할 변수들은 다음과 같다.

|  |  |  |  |
| --- | --- | --- | --- |
| 변수명 | 설명 | 변수명 | 설명 |
| Country name | 국가명 | Regional Indiator | 대륙이 속한 지역  (서유럽, 동유럽 등) |
| Ladder Score | 행복도 산출 점수 | Logged GDP Per capita | 1인당 GDP |
| Social Support | 삶에 문제가 생겼을 때 도움을 줄 수 있는 사람이 있는지에 대한 여부를 점수화 | Healthy Life Expectancy | 기대 건강 수명  (건강하게 살 수 있는 나이) |
| Freedom | 피조사자가 느끼는 자유도를 점수화 | Generosity | 각 국가의 기부문화를 점수화  (점수가 낮으면 사회분위기가 각박함을 의미함) |
| Perception of corruption | 피조사자가 느끼는 부패정도를 점수화 |  |  |

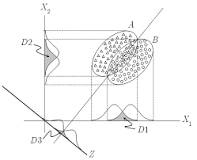
< 표 1: 데이터 변수 설명>

본 데이터는 <표 1>과 같이, 2021년 각 국가별로 위의 변수를 조사하여 발표한 데이터이다. 이 때, 위의 Ladder Score 변수는 타 변수의 값을 공식에 대입하여 산출한 값이다.

**1.4 분석 방법**

본 연구에서는 다음의 몇 가지 다변량 통계학적 분석기법을 사용하여 분석할 예정이다.

* 판별분석 : 행복도가 높은 나라와 그렇지 못한 나라의 차이점을 분석하는데 사용 / 대륙별 차이점을 알아보는데 사용할 예정이다.
* 주성분 분석 : 데이터 수가 많지가 않기 때문에, 판별분석을 시행하기 이전에 차원축소를 위해 사용한다.
* 호텔링의 T2 통계량을 활용한 검정 : 밝혀낸 군집 내에서 Outlier의 존재여부를 살피기 위해 검정을 시행한다.
* 수업이 진행됨에 따라 배우는 여러 분석기법도 추가적으로 참고하여 분석할 예정이다.

판별분석은, 몇 개의 알려진 집단으로부터 그룹의 특성을 나타내고 구별해주는 판별함수 또는 판별규칙을 만들고, 결정된 판별함수를 이용하여 새로운 개체를 어느 집단에 속하는지를 판별하여 분류하는 다변량 기법이다. 이 때, 표준화된 변수를 이용하여 산출한 판별함수의 계수는 각 변수의 중요성의 지표가 되므로, 행복도 별 / 대륙 별 변수들의 차이를 분석하는 데 유용할 것이라고 판단하였다. 또한, 행복도가 높지도 낮지도 않은 중간권의 국가들에 대하여 판별분석을 통해 판별함으로써, 이들 국가가 어느 집단의 특성에 더 가까운지, 혹은 이들 중 행복도가 높아질 잠재력이 있는 국가가 어느 것인지에 대한 추측 또한 가능할 것이라고 생각하였다.

<그림 1: 판별분석 설명 자료>

주성분 분석은 다변량 데이터를 이해하기 쉽게 만드는 다변량 기법으로, 원래 변수들의 선형결합을 이용하여 주어진 데이터를 원 변수 개수보다 적은 수로 원래 정보와 근접하게 표현하는 것이다. 보통 이는 다변량 변수의 차원 축소, 데이터 해석등의 목적으로 주로 사용된다.

호텔링의 T2 은 일변량변수의 T 통계량의 다변수에서의 일반화로, 다변량 벡터 간의 거리를 계산하여 모평균검정을 시행하는데 사용할 수 있다.

**1.5 결과 활용 및 기대효과**

본 연구를 통하여 다음의 몇가지 결과를 얻을 수 있다.

1. 세계 각국의 Whr 데이터에 대한 주성분.

2. 행복도가 높은 나라와 작은 나라를 구분하는 판별함수. 이를 통해 행복도의 대소에 큰 영향을 미치는 변수를 확인 할 수 있다.

3. 2에서 구한 판별함수를 이용하여 행복도가 중간인 나라 중 어느나라가 행복도가 높아지기 더 유리한 조건인지 판별할 수 있음

4. Whr 데이터에 대한 세계 지역 판별함수

5. 소속된 지역의 특징과 거리가 있는 국가 확인 등

위의 결과들을 활용하여 기대할 수 있는 효과는 다음과 같다.

* 국가의 행복도를 높이기 위한 의사결정에 활용
* 이민/유학 준비자들의 의사결정에 활용

1. **본론**
   1. **분석 방법 소개**

본 연구는 크게 기본 전처리, 행복도의 대소에 대한 판별분석, 국가가 속한 지역에 관한 판별분석의 세 가지 부분으로 나누어 수행하였다.

1. 데이터 기본 전처리

* 변수 선택
* 주성분 분석을 위한 데이터 표준화
* 주성분 분석과 해석

1. 행복도 대소에 대한 판별 분석

* 행복도 상위 25개, 하위 25개에 대한 판별 분석
* 상/하위 50개 데이터에 속하지 않는 99개의 데이터를 Test set으로 사용, 전 단계에서 구한 판별함수를 이용한 판별

1. 국가가 속한 지역에 관한 판별분석

* 원 데이터의 Regional\_indicator(속한 지역)에 대한 판별분석
* 전 단계의 판별분석의 결과를 이용하여 비슷한 특징을 가지는 지역을 병합.
* 데이터의 다변량 정규성 검토
* 위의 병합 처리에 대한 분산분석
* 각 군집별로 T2 통계량 이용하여 Outlier 탐색
  1. **데이터 분석 및 결과 설명**

1. **데이터 기본 전처리**

분석 대상 변수는 1인당 GDP, 주변인들의지원, 건강 기대수명, 선택의 자유, 기부문화, 부패인식의 6가지로 정하였다.

이후 이 6가지 변수들에 대하여, 주성분 분석을 수행하였다. 이를 수행하기 전에, 위의 6가지 변수들에 대한 표준화를 진행하였다.

이를 수행한 SAS 코드와 결과는 다음과 같다.

/\*주성분분석 전 데이터표준화\*/

**proc** **standard** data = whr\_preprocessed out = whr\_std mean = **0** std= **1**;

var Logged\_GDP\_per\_capita -- Perceptions\_of\_corruption;

**run**;

/\*주성분분석\*/

**proc** **princomp** data = whr\_std out = pca covariance;

var Logged\_GDP\_per\_capita -- Perceptions\_of\_corruption;

**run**;

위의 코드를 통해 주성분 분석을 수행한 결과는 다음과 같다.

테이블이(가) 표시된 사진

자동 생성된 설명

<표 1 : 주성분 분석 수행 결과>

위의 결과를 해석하면, 네 개의 주성분으로 총 분산의 93 % 이상을 설명할 수 있다. 위로부터 구한 주성분을 해석한 결과는 다음과 같다.

* Prin1 : GDP, 사회적 지원 ,건강기대수명, 자유도가 골고루 높은 선진국여부
* Prin2 : 기부문화가 발달한 국가 여부
* Prin3 : 커뮤니티가 발달했고, 부패한 국가인지의 여부
* Prin4 : 자유롭지 못한 독재국가 여부

1. **행복도 대소에 대한 판별분석**

이전 단계에서 구한 주성분을 바탕으로, 국가의 행복도 대소에 관한 판별분석을 진행하였다. 먼저, 행복도 상위 25개국과 하위 25개국을 추출하여 이들에게 각각 class 1과 0을 부여하였다. 그리고 구해진 주성분을 바탕으로 부여된 class 에 따른 판별분석을 수행하였다. 판별분석 이전에, 구해진 주성분을 한번 더 표준화 시키는 작업을 수행하였다. 이를 수행한 SAS 코드는 다음과 같다.

/\*1. 상위 25개국과 하위 25개국을 나누어 class 부여한다.\*/

**proc** **standard** data = pca out = pca\_std mean = **0** std = **1**;

var Prin1--Prin6;

**run**;

**data** pca\_ranked;

set pca\_std;

row\_number = \_N\_;

**run**;

**data** pca\_classed;

set pca\_ranked;

if row\_number <= **25** then happy = **1**;

if row\_number >= **125** then happy = **0**;

if row\_number>=**26** and row\_number <**125** then happy = **.**;

**run**;

**data** pca\_classed\_for\_discrim;

set pca\_classed;

where happy = **1** or happy = **0**;

**run**;

**data** pca\_classed\_for\_test;

set pca\_classed;

where happy = **.**;

**run**;

/\*2.판별분석\*/

**proc** **discrim** data = pca\_classed\_for\_discrim simple listerr pool = yes;

class happy;

var Prin1--Prin4;

priors prop;

**run**;

참고로, 판별분석 수행 시, 선형 판별함수를 산출하기 위하여 두 집단의 공분산은 같다고 가정하고 분석을 수행하였다.

이를 통해 산출한 결과는 표 2에 나타나 있다.

표 2 : 얻어진 선형 판별함수 / 표 3 : 판별분석 분류 결과

테이블이(가) 표시된 사진

자동 생성된 설명 테이블이(가) 표시된 사진

자동 생성된 설명

표 2의 결과에 따르면, 구해진 판별함수는 다음과 같다.

Y = (4.83-(-4.42))\*Prin1 + (-0.14-0.77)\*Prin2 +(-0.23-(-0.56))\*Prin3 + (1.58-(-1.00)\*Prin4

**=** 9.25\*Prin1 -0.91\*Prin2 + 0.33\*Prin3 + 2.58\*Prin4

이로부터 행복도의 대소에 가장 영향을 많이 주는 요인은 Prin1(선진국여부), Prin4(독재국가여부), Prin2, Prin3순임을 알 수 있다. (표준화된 판별함수 이므로)

표 3의 분류표를 보면, 위의 판별함수는 총 50개 데이터 중 49개의 데이터를 정확히 분류했음을 알 수 있다.

구해진 판별함수를 이용하여, 총 149개국 중 상/하위 25개국을 제외한 나머지 99개국에 대해 판별함수를 적용하여 보았다. 판별함수는 99개국 중 52개국을 행복도가 높을 것으로 예측했고, 나머지 47개국을 행복도가 낮을 것으로 예측하였다.

테스트 결과는 다음과 같다.

* 예상보다 행복도가 높은 국가 (판별함수는 행복도가 낮을 것으로 예측했으나 실제로는 상대적으로 높은 국가)

: 과테말라, 자메이카, 엘살바도르, 온두라스 ,페루, 필리핀, 코소보

* 예상보다 행복도가 낮은 국가 (판별함수는 행복도가 높을 것으로 예측했으나 실제로는 상대적으로 낮은 국가)

: 이란, 레바논, 터키, 투르크메니스탄, 이란, 우크라이나, 알제리 , 베네수엘라

예상보다 행복도가 높은 국가들의 경우, 조사된 변수들은 행복도가 낮은 나라에 가까운 경향을 보이나, 상대적으로 행복도가 높은 것으로 보아 주성분만으로는 설명이 안되는 특정요인이 있다고 생각할 수 있다. 특이한 점은, 이 국가들의 상당 수가 중남미 지역에 위치한 국가들이라는 점이다. 이를 통해 볼 때, 중남미 지역 국가들이 공유하는 특징이 이들 국가의 예상보다 높은 행복도에 기여하고 있다고 볼 수 있다.

반면 예상보다 행복도가 낮은 국가들의 경우, 주로 중동/아랍 지역에 위치해있는 것을 볼 때, 이들 국가들이 공유하는 특징이 이들 국가의 예상보다 낮은 행복도의 원인이 된다고 볼 수 있다. 예상보다 행복도가 낮은 국가들은, 조사변수의 값이 행복도가 높은 국가의 그룹에 속해있기 때문에, 낮은 행복도의 원인을 찾고 이를 보완하면 빠른 시일내에 행복도가 올라갈 수 있을 것이라고 예측된다.

1. **국가가 속한 지역에 대한 판별분석**
2. **각 대륙별 판별분석과 행복도별 군집화**

이번에는 앞 단계에서 구한 표준화된 주성분을 바탕으로, 각 국가가 속한 대륙에 대한 판별분석을 시행하였다. 판별결과의 일부는 다음과 같다.

테이블이(가) 표시된 사진

자동 생성된 설명

표 3 : 대륙별 판별분석 결과 일부

판결분석 결과를 요약하면 다음과 같다.

* 분류가 잘 된 지역: Central and Eastern Europe(12/17 옳게 분류), Latin America and Caribbean(15/20), Western Europe(17/21), sub-saharan Africa(31/36)
* 특정 대륙의 경향을 띄는 지역 : middle east and north Africa(9/17), north America and Anz(모두 western europe로 분류), south-asia (5/7이 sub-saharan africa로 판별)
* 그 외 지역 : commonwealth of independent / eastasia (데이터 적음), southeast asia(판별결과 혼재)

이 결과로부터 알 수 있는 특이한 점은 북아메리카 지역의 특성은 서유럽지역과 거의 비슷하고, 남아시아 지역은 사하라 이남 아프리카 지역과 특성을 공유한다는 것이다.

위의 결과를 참고하여, 위치 지역이 아닌 행복도 기준으로 다시 세계를 분류한 후 분석을 시행하기로 하였다. 위의 결과에 따른 분류결과는 다음과 같다.

<1. 중-동 유럽형 2. 남아메리카형 3. 서유럽형 4. 사하라 이남 아프리카형 5. 북아프리카형>

1. **행복도 별 군집화에 대한 분산분석**

행복도별 군집 각각에 대한 분석에 들어가기에 앞서, 위의 분류가 통계적으로 유의미한지를 확인하기 위해 분산분석을 시행하겠다. 분산분석 시행에 앞서, 다변량 데이터들이 다변량정규분포를 따르는지 우선적으로 확인하기로 했다. 이를 위해 활용한 방법론은 다음과 같다.

* 각각 다변량벡터 데이터에 대하여, 각각 데이터가 다변량 정규분포를 따른다면, 데이터의 평균으로부터의 마할라노비스거리의 제곱은 자유도가 P=4(변수 수)인 카이제곱 분포를 따른다.
* 자유도가 P=4인 카이제곱분포와 각 데이터의 마할라노비스 거리 간의 QQ-Plot을 그려서, 이 그래프가 일직선과 비슷한 형태를 띈다면 다변량 데이터는 다변량 정규분포를 따른다고 볼 수 있다.

이 과정을 수행한 sas 코드와 qq-plot은 다음과 같다.

**proc** **iml**;

use X;

read all into x;

mean = mean(x);

cov = cov(x);

md = (x - J(**149**,**1**)\*mean) \* inv(cov) \* (x - J(**149**,**1**)\*mean)`;

md2 =diag(md)\*J(**149**,**1**);

call sort(md2, **1**);

s = (T(**1**:**149**) - **0.375**) / (**149** + **0.25**);

chisqQuant = quantile("ChiSquare", s, ncol(X));

call scatter(md2, chisqQuant);

**run**;

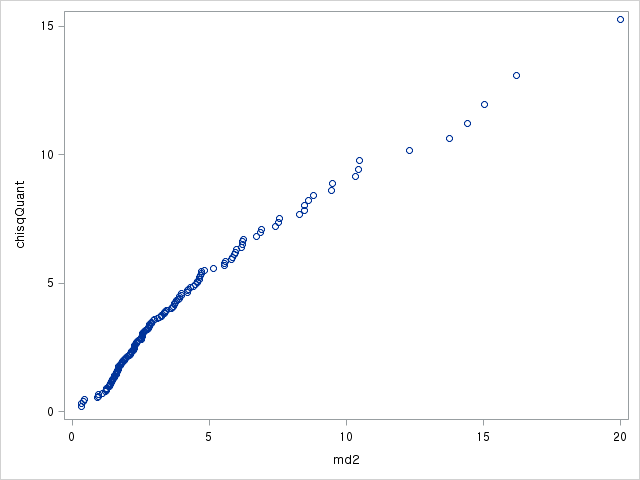


그림 2 : 카이제곱 QQ-Plot

그려진 QQ-Plot을 보면, 다변량데이터가 다변량정규분포에 가깝게 분포되어있다는 것을 확인할 수 있다. 따라서, 논리적 결함 없이 다변량분산분석이 가능하다. 이를 수행한 결과는 다음과 같다.

**proc** **anova** data = pca\_std\_clustered;

class cluster;

model Prin1 -- Prin4 = cluster;

manova h =cluster ;

테이블이(가) 표시된 사진

자동 생성된 설명테이블이(가) 표시된 사진

자동 생성된 설명

**<표 4/ 5: 다변량 데이터에 대한 분산분석 결과>**

위의 결과를 보면, p-value가 0.0005가량으로, 행복도 군집에 따른 국가의 분류가 유의미하다는 사실을 알 수 있다.

1. 텍스트이(가) 표시된 사진

   자동 생성된 설명**군집별 모평균 벡터의 특징**

* 중부 – 동부 유럽은 기부문화가 발달하지 못했고, 커뮤니티가 발달한 특성을 띈다
* 북아프리카 지역은 기부분화가 발달하지 못했고, 자유도가 상대적으로 낮다
* 남아프리카지역은 경제적으로 부유하지 못하다
* 남아메리카 지역은 자유도가 상대적으로 높다
* 서유럽 지역은 부유한 선진국이 많이 위치했으며, 자유도가 상대적으로 낮다.

1. **군집 내 이상치에 대한 분석.**

행복도 군집에 따른 분류가 유의미하다는 것을 확인했으니, 이제 각 집단에 대해서 어느 국가가 그 집단의 특성에서 엇나가 있는지에 대한 분석을 수행하겠다. 이를 수행하기 위하여, 호텔링의 T2 통계량(마할라노비스 거리)를 활용하였다. 이에 따르면, 각 군집 내 평균으로부터 데이터 까지의 마할라노비스 거리는 자유도가 4인 카이제곱 분포를 따르며, 유의수준 a = 0.05수준에서 검정하기 위하여, 구한 마할라노비스 거리가 자유도가 4인 카이제곱분포의 95분위수(9.49)보다 클 경우 이상치로 판단하였다.

/\*proc iml 을 이용한 호텔링 통계량(마할라노비스 거리) 이용한 검정\*/

/\*1. 중부유럽\*/

**proc** **iml**;

use ceEurope;

read all into X;

uvec = J(**1**, nrow(X))\*X/nrow(X);

u = J(nrow(X), **1**)\*uvec;

cov = (X`\*(I(nrow(X)) - J(nrow(X))/nrow(X))\*X) / (nrow(X) -**1**);

T2 = diag((X-u)\*inv(cov)\*(X-u)`);

print(u);

print(cov);

print(T2);

**run**;

산출값 예시는 다음과 같다. (중-동 유럽 예시)

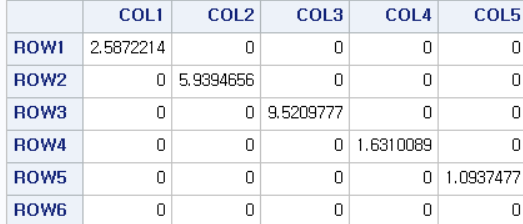


표 6 : 호텔링 통계량 산출 예시(여기서 3번 데이터는 이상치로 판단할 수 있다.)

위와 같이 분석한 결과를 정리하면 다음과 같다.

* 중부-동부 유럽 : 이상치 값 3(코소보), 6번(에스토니아) 데이터
* 북아프리카 : 11번(이란) 데이터
* 남아프리카 : 7(아프가니스탄), 15(감비아), 42(르완다)번 데이터
* 서유럽 : 23(그리스), 24(포르투갈)번 데이터
* 남아메리카 : 20(아이티)번 데이터

이제, 각 이상치에 대한 분석을 시행하겠다.

* 중부-동부 유럽 : 코소보, 에스토니아

우선, 중부-동부 유럽의 모평균 벡터를 통해 볼 때,



표 : 7 – 중부/동부 유럽형의 모평균벡터

텍스트이(가) 표시된 사진

자동 생성된 설명

표 8 : 코소보, 에스토니아의 주성분 값

중부유럽은 1. 커뮤니티가 발달하고 부패(주성분 3) 2. 기부문화가 미성숙(주성분2)의 특성을 띈다. 코소보의 경우, 주성분 2와 3이 높은 것으로 나타나며, 에스토니아의 경우 주성분 1이 높고 주성분 4가 낮은 것으로 나타난다. 즉, 에스토니아의 경우 중부-동부 유럽 군집보다 서유럽 군집과 비슷한 특성을 띄며, 코소보의 경우 속한 군집의 평균보다 더 열악한 환경을 가지고 있음을 알 수 있다. 사회-문화적 상황과 이 결과를 결부해서 본다면, 에스토니아의 경우 발트3국의 최북단의 위치하여 서-북유럽과 지리적으로 매우 가깝고, 타 발트3국에 비해서 북유럽으로부터 사회-문화적 영향을 많이 받은 점이 데이터에도 반영 된 것으로 보인다.

코소보의 경우, 위치적으로는 발칸반도의 한복판에 위치하고 있으나, 현재 알바니아-세르비아 간 분쟁지역이라는 특수성이 데이터에도 영향을 미친 것으로 판단된다.

* 북아프리카 : 이란

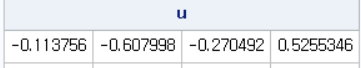


표 9 : 북아프리카 형의 모평균벡터



표 10 : 이란의 주성분벡터

이란의 경우, 주성분 4(독재국가)의 특성이 북아프리카 평균에 비해서 매우 뚜렷하게 나타난다. 이러한 특성 때문에, 이란은 이상치로 나타난 것으로 보이며, 이는 현재 이란의 사회적 상황과도 일치한다.

* 남아프리카 : 아프가니스탄, 감비아, 르완다

아프가니스탄과 감비아의 경우, 독재국가 특성이 남아프리카 평균보다 매우 크게 나타나며, 아프가니스탄의 경우, 주성분 1을 볼 때 경제상황이 유독 좋지 않은 것으로 보이며, 기부문화수준, 커뮤니티 발달수준(주성분 2,3) 도 크게 낮다. 감비아의 경우, 기부문화가 남아프리카 평균에 비해 발달해있으며, 커뮤니티 문화 역시 발달 한 것으로 판단된다. 르완다의 경우, 기부문화가 발달했으며, 커뮤니티 문화는 매우 발달하지 못한 양상을 띈다.

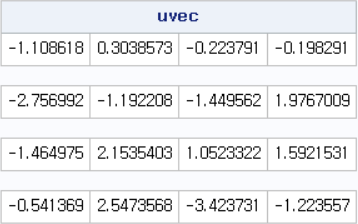


표 11 : 남아프리카 모평균 벡터와 아프가니스탄, 감비아, 우간다의 주성분벡터

* 서유럽 : 그리스, 포르투갈

테이블이(가) 표시된 사진

자동 생성된 설명

표 12 : 서유럽 모평균 벡터와 그리스, 포르투갈의 주성분벡터

서유럽에 선진국들이 몰려있는 사실을 반영하듯, 서유럽은 평균적으로 선진국이 많고(주성분 1), 다소 개인주의적 성향을 띈다(주성분 3).

그리스와 포르투갈의 공통점은 일단 두 나라 모두 선진국기준에 해당할 만큼 부유한 국가들은 아니라는 것이다. 실제로, 그리스의 1인당 GDP는 18000달러수준, 포르투갈의 경우 24000수준으로 대한민국보다도 상당히 낮은 수준에 있다. 또한 지정학적으로 두 나라모두 서유럽의 변방에 위치해있다는 것 역시 눈여겨 볼만 하다.

경제적 요인 뿐만 아니라, 그리스/포르투갈 모두 두드러지게 기부문화가 발달하지 못한 것을 볼 수 있다.(주성분 2)

* 남아메리카 : 아이티



표 13: 남아메리카의 모평균 벡터와 아이티의 주성분 벡터

아이티의 경우, 경제적 수준이 매우 낮으며(주성분 1), 기부문화가 발달해있고, 독재국가라는 점이 남아메리카의 평균적인 특성과 이들을 분류하는 가장 큰 요인이다.

1. **결론**

**3.1 분석 결과 요약**

본 연구에서는 알아보고자 한 것은 크게 두 가지였다.

1. 행복도가 높은 국가와 낮은 국가를 판별하는 요소는 무엇인가?
2. 국가가 위치한 지역별로 관찰변수는 어떠한 양상을 띄는가?

먼저, 첫 번째 주제에서는,

Y **=** 9.25\*Prin1 -0.91\*Prin2 + 0.33\*Prin3 + 2.58\*Prin4

의 선형판별함수를 도출했으며, Prin1, Prin4, Prin2, Prin3 순으로 영향력을 가지는 것으로 나타났다. 즉, 행복도 상위 국과 하위 국가를 구분하는 데에는 그 나라가 선진국인 여부가 가장 큰 영향을 끼치며, 그 다음으로 자유도가 부족한 독재국가인지 여부가 큰 영향을 끼쳤다. 위에서 도출한 판별함수를 사용하여 테스트 국가들을 분별해본 결과, 중남미 지역의 국가들 중 상당 수는 예상보다 실제 행복도가 높음을 알 수 있었고, 반면 중동/아랍 지역의 다수 국가들이 판별함수로 얻은 행복도보다 실제 행복도가 낮은 편에 속한다는 것을 알 수 있었다.

그 다음으로, 국가 위치 지역과 행복도의 관련성에 대해 조사한 두 번째 주제에서는, 판별분석 결과와 다변량 분산분석 결과에 따라 세계지역을 행복도에 따라 크게 5지역으로 분류할 수 있음을 보였으며, 각 지역의 평균적인 특성과, 호텔링 통계량을 이용하여 그 집단 내에 존재한 이상치들을 탐색한 후 관찰하였다.

관찰 결과, 군집 내에서 나타난 이상치들은 대부분 일반적이지 않은 사회, 문화적 상황을 겪고 있거나(이란, 코소보, 아프가니스탄) 위치적으로 해당 분류 지역의 외곽에 위치하는 등의 특성을 가지고 있었다. (에스토니아, 그리스, 포르투갈) 이는 연구 초반에서, 지정학적/사회문화적 요인이 행복도에 큰 영향을 끼칠 것이라고 예상했던 바와 상통한다.

**3.2 분석의 장점 및 한계점 설명**

본 분석은 신뢰도가 있는 데이터를 사용하여, 주로 판별기법을 이용하여 데이터를 분석하였다. 이 과정에서, 분석기법의 객관성을 위해서 필요한 가정들을 만족시키기 위해 노력하였다. 주성분기법/판별분석 기법을 하기 이전의 데이터 정규화 작업과, 다변량 분산분석을 시행하기 이전에 수행한 다변량 정규분포 검정이 그 예시이다.

또한, 도출한 판별함수를 모형 학습에 사용되지 않은 테스트 데이터에 활용함으로써, 도출한 모델의 활용도를 높였다. 그 결과 테스트 데이터들에 대한 직관을 도출할 수 있었다.

본 분석의 한계점으로는 몇 가지가 있는데, 첫 번째로, 분석 초반에 도출한 주성분의 해석에 있어서 다른 해석의 여지가 있다. Prin1과 Prin4의 해석은 비교적 간단하지만, Prin2 와 Prin3의 경우 다른 해석의 여지가 있었기 때문에, 보다 깊은 논의가 어려웠다.

또한, 지역에 따른 행복도의 분석에서 탐색한 이상치들에 대하여, 그들의 주성분 벡터와 평균벡터의 대소 비교에서 논의가 멈췄고, 더 깊은 탐색이나 통계적 검정을 하지 않아, 다소 분석의 활용도가 떨어진다는 단점이 있다.

**3.3 추가 연구사항 제안**

본 분석의 아이디어를 바탕으로, 판별분석 이외의 다양한 방법론들을 통하여 국가의 특성과 행복도에 대한 분석을 진행한 후, 정책적인 연구를 거쳐 해당 국가에 행복도를 높이기 위한 자세한 가이드라인을 제시할 수 있다면, 활용도가 매우 높고 가치있는 연구가 될 것이라 생각한다.

**#. 참고문헌**

[1]데이터 출처 : https://www.kaggle.com/datasets/unsdsn/world-happiness

[2]Methods of Multivariate Analysis 3th ed. – Alvin. C. Rencher