**다변량통계학 학기말과제 예비보고서**

<World Happiness Report를 활용한 국가 군집별 특징분석>

담당교수 : 김규성 교수님

제출일자 : 2022.11.28

**2019580027 정우진**

목차

**1. 서론3**

1.1. 연구 목적3

1.2. 문헌 연구 3

1.3. 데이터 설명 4

1.4 분석 방법5

**2. 예상 결과6**

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?
* 분석 결과에 따라 보이는 특이점에 대한 추가적인 분석

**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 | Ladder score in dystopia | Dystopia는 Utopia의 반대를 의미하며, 상상속의 최악의 국가를 의미. 이 변수는 피조사자가 상상하는 최악 국가의 행복도 산출점수를 조사한 것. 자신의 현 상황에 대한 상대적인 인식을 의미함 |

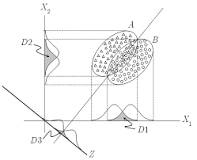
< 표 1: 데이터 변수 설명>

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.
* 정준상관분석 : 행복도 간 / 대륙 간 벡터의 분포의 상관관계를 알아보는데 사용할 예정이다.
* 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>

정준상관분석은, 다차원 벡터집단 간의 선형 상관성을 측정하는 분석법이다. 이 방법은 각 집단 별 변수들의 선형 결합 간의 상관계수를 계산하여 두 집단간의 상관성을 측정한다. 본 기법을 사용하여, 행복도 간 / 대륙 간 벡터들의 연관성 양상을 보고자 한다.

1. **예상되는 결과**
2. 행복도에 따라 나눈 집단간의 분석

두 집단의 변수들은 전반적으로 큰 차이를 보일 것으로 예상된다. 집단 분류에 사용되는 행복도는 분석대상 변수들을 공식에 대입하여 점수화 한 것이기에 이는 자명하다.

구해진 판별함수의 계수값을 이용하여 변수 별 중요도를 분석해보면, 1인당 GDP가 가장 큰 영향력을 끼칠 것으로 예상된다. 또한 정준상관분석을 통한 다변수 벡터간 상관관계는 크게 있지 않을 것으로 예상된다. 예상에 따라 경제적 수준이 가장 큰 영향력을 가진다고 생각하면, 두 집단간엔 큰 경제적 차이가 있을 것이고, 경제적 차이가 크게 차이나는 집단 간에 다른 변수들이 큰 상관관계를 가진다고 생각하기 어렵기 때문이다.

1. 지역별로 나눈 집단간의 분석

지역집단 간, 특히 부유한 나라들이 모여있는 서유럽/북아메리카의 경우 아프리카와 같은 상대적으로 가난한 지역과 큰 차이가 있을 것으로 예상된다. 또한 비슷한 경제적상황을 가진 지역들 사이에도 유의미한 차이가 있을 것으로 예상된다. 지정학적 요인에 따라 변수간에 두드러지는 차이가 있을 것으로 생각하기 때문이다.

정준상관분석을 통한 다변수 벡터간 상관관계의 경우, 1의 경우와 마찬가지로 크지 않을 것으로 예상된다.

이와 같은 예상에 기초하여 분석을 하고, 결과가 예상과 다르다면 왜 그러한지에 대해서도 분석할 예정이다.

**#. 참고문헌**

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

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