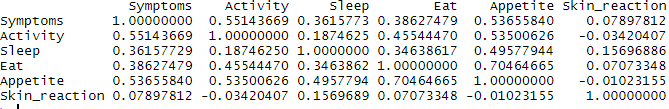
1. 다음 Table1\_7.txt 파일은 환자들이 방사능치료를 받을 때 다음 6개의 항목에 대해 평가한 평가점수들을 나타낸다. Symptoms는 목이 쓰림, 현기증과 같은 증상의 수를, Activity는 1-5 리커트 척도인 활동의 양을, Sleep은 1-5 리커트 척도인 잠의 양을, Eat는 1-3 리커트 척도인 섭취한 음식의 양을, Appetite는 1-5 척도인 식욕을, Skin reaction은 0-3 리커트 척도인 피부 반응을 나타낸다.

a) 각 변수들의 평균과 상관계수 행렬을 구하시오.

Symptoms의 평균 : 3.542 Activity의 평균 : 1.809 Sleep의 평균 : 2.138

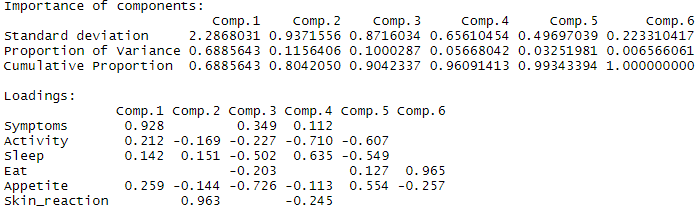
Eat의 평균 : 2.209 Appetite의 평균 : 2.575 Skin reaction의 평균 : 1.276

상관계수 행렬

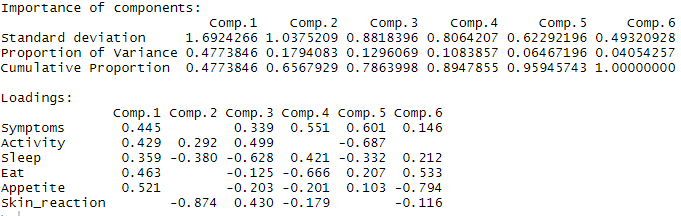


b) 공분산과 상관계수를 이용한 주성분 분석을 수행하고 어떤 방식이 나은지 평가하시오.

공분산을 이용한 주성분 분석

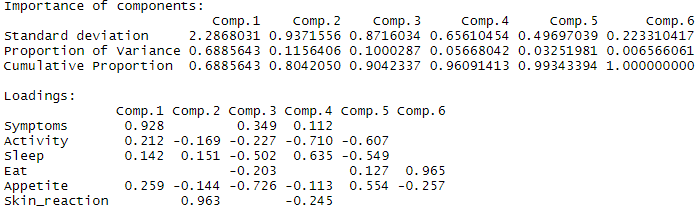


상관계수를 이용한 주성분 분석



공분산을 이용한 주성분 분석이 상관계수를 이용한 주성분 분석보다 총분산의 누적비율이 더 높다. 또한 변수들의 범위나 척도가 매우 다르지 않고 유사하여 분산을 보존하기 위해 공분산행렬을 사용하는 것이 낫다.

c) b)에서 선택한 방식에 의해 주성분 분석을 수행하였을 때 첫번째와 두번째 주성분에 의해 설명되는 분산의 비율은 얼마인가? 80.4%



d) 적절한 주성분 개수를 제안하시오. 2개

e) 가능하다면 첫번째, 두번째, 세번째 주성분이 나타내는 의미를 설명하시오.

첫번째 주성분은 symptoms를 가장 잘 설명하는 속성이다. 첫번째 주성분의 점수가 높으면 목이 쓰리거나 현기증 증상이 자주 나타나는 거다.

두번째 주성분은 skin\_reaction(피부 반응)을 가장 잘 설명하는 속성이다. 두번째 주성분의 점수가 높으면 피부 반응이 잘 나타나는 것이다.

세번째 주성분은 appetite(식욕)과 sleep(잠의 양)을 잘 설명하는 속성이다. 세번째 주성분의 점수가 낮으면 식욕과 잠의 양이 많은 것이다.

radio <- read.table("C:\\Table1\_7.txt")

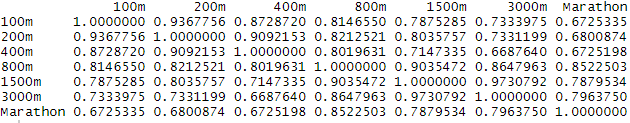
colnames(radio) = c("Symptoms", "Activity", "Sleep", "Eat", "Appetite", "Skin\_reaction")

head(radio)

2. 다음 Table1\_9.txt 파일은 각 나라의 여자 육상선수들의 최고기록들을을 나타낸다. 예로서 아르헨티나(ARG)의 100m 여자 최고기록은 11.57초를 나타낸다.

a) 표본 상관행렬을 구하고 고유값과 고유벡터를 구하시오.

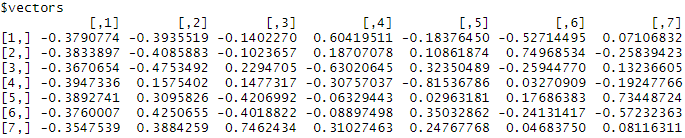
표본상관행렬



고유값

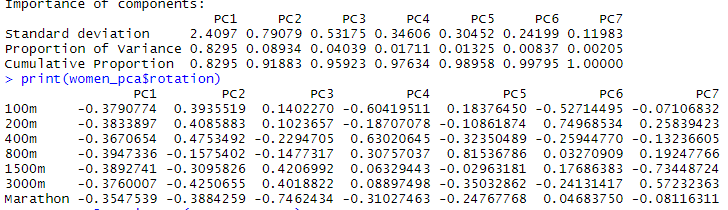


고유벡터



b) 상관행렬을 이용하여 주성분분석을 수행하고 적절한 주성분개수를 제안하시오.

1개



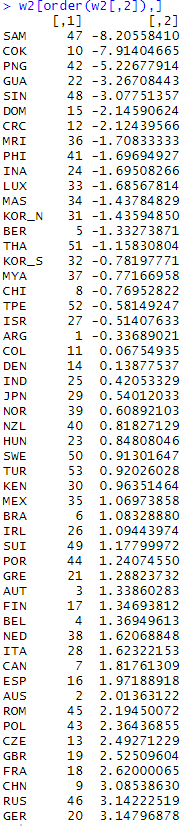
c) b)에서 얻은 처음 두개의 주성분을 해석하시오.

첫번째 주성분은 모든 종목의 점수를 잘 설명하고 총분산의 누적비율이 82%이다.

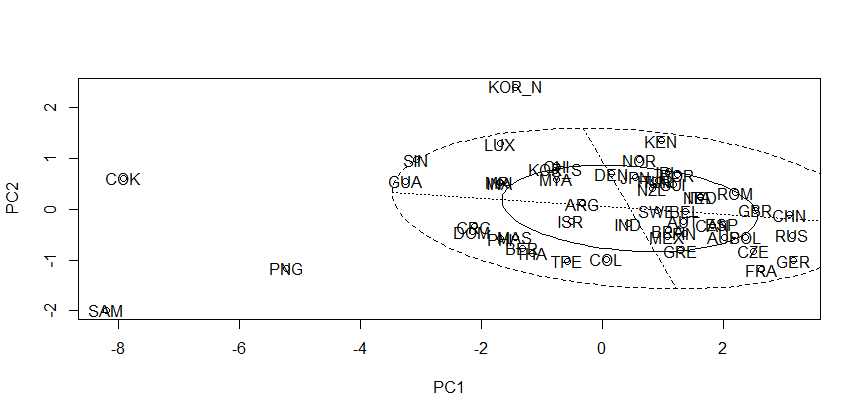
두번째 주성분은 점수가 절대값이 큰 양수를 가지면 100m, 200 ,400m 종목의 기록이 높고 1500m, 300m, marathon 의 기록이 낮다. 점수가 절대값이 큰 음수를 가지면 1500m, 300m, marathon의 기록이 높고 100m, 200 ,400m 종목의 기록이 낮다.

d) 첫번째 주성분 점수가 낮은 순서대로 국가들을 정렬하시오. 이 순서가 운동을 잘하는 나라라고 상식적으로 생각되는 순서와 일치한다고 생각하는가?

아니요. 첫번째 주성분의 점수가 높은 나라가 운동을 잘하는 나라이다.

. 

e) 첫번째 주성분점수와 두번째 주성분 점수의 산점도를 그리시오, 그리고 각 위치에 나라의 이름이 기록되게 하시오.



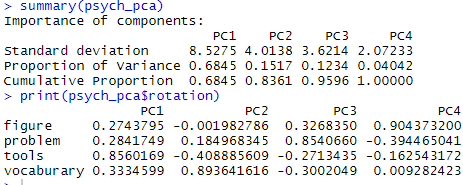
women <- read.table("C:\\Table1\_9.txt”)

colnames(women) = c("Country", "100m", "200m", "400m", "800m",

"1500m", "3000m", "Marathon")

3. 다음 psychiatric.txt 파일은 figure, problem, tools, vocaburary의 인식 능력에 대한 심리검사 자료로서 성별(sex)는 남성은 1, 여성은 2로 표현되어 있다.

a) 전체 데이터에 대해 공분산행렬을 이용하여 주성분분석을 수행하고 적절한 주성분개수를 제안하시오. 2개 (총분산의 누적비율이 제2주성분이 83%이므로)



b) a)의 첫번째 주성분을 해석하시오.

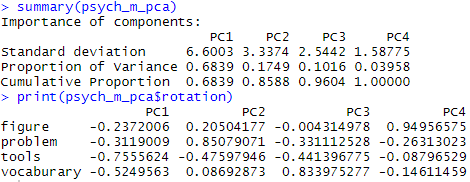
첫번째 주성분은 tools를 가장 잘 설명하는 속성이고 첫번째 주성분 점수가 높으면 tools의 인식 능력이 높은 것이다.

c) a)의 두번째 주성분을 해석하시오.

두번째 주성분은 vocaburary를 가장 잘 설명하는 속성이고 두번째 주성분의 점수가 높으면 vocaburary의 인식 능력이 높은 것이다.

d) 남성 데이터에 대해 공분산행렬을 이용하여 주성분분석을 수행하고 적절한 주성분개수를 제안하시오.

2개 (총분산의 누적비율이 85%이므로)



e) d)의 첫번째 주성분을 해석하시오.

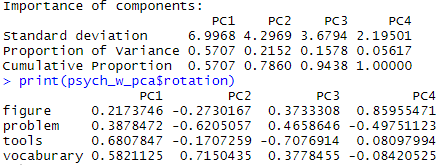
첫번째 주성분은 tools와 vocaburary의 인식기능을 잘 설명해주고 있다. 첫번째 주성분 점수가 높으면 tools와 vocaburary의 인식기능이 낮다는 것을 보여주고 있다.

f) d)의 두번째 주성분을 해석하시오.

두번째 주성분은 problem을 잘 설명해주고 있다. 두번째 주성분 점수가 높으면 problem의 인식기능이 높다는 것을 보여주고 있다.

g) 여성 데이터에 대해 공분산행렬을 이용하여 주성분분석을 수행하고 적절한 주성분개수를 제안하시오.

3개 (총분산의 누적비율이 94%이므로)



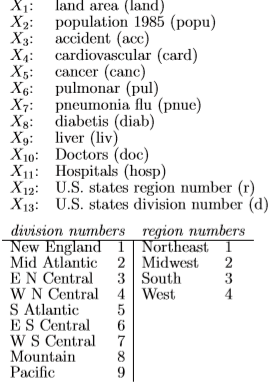
h) g)의 첫번째 주성분을 해석하시오.

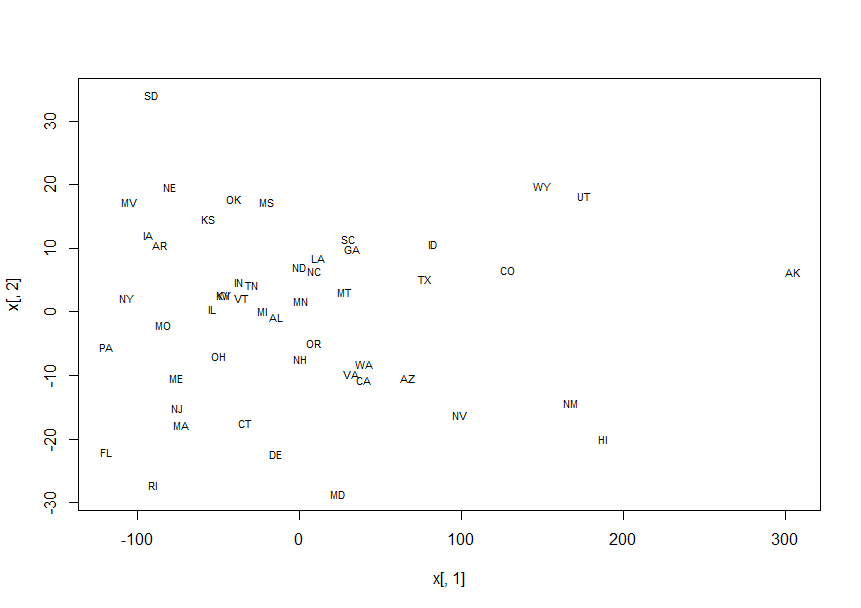
첫번째 주성분은 tools와 vocaburary를 잘 설명해주고 있다. 첫번째 주성분의 점수가 높으면 tools와 vocaburary의 인식기능이 높다는 것이다.

i) g)의 두번째 주성분을 해석하시오.

두번째 주성분은 vocaburary와 problem을 잘 설명해주고있다. 두번째 주성분 점수가 절대값이 큰 양수이면 vocaburary의 인식기능이 높고 problem의 인식기능은 낮다. 두번쨰 주성분 점수가 절대값이 큰 음수이면 problem의 인식기능은 높고 vocaburary의 인식기능은 낮다.

4. UShealth.txt 파일은 1985년에 미국의 50개 주에서 7개의 범주(X3 – X9)에 따라 분류된 사망자수(10만명당)들을 나타낸다. 7개의 범주(X3 – X9)에 따라 다차원척도법을 시행하고 2차원 공간에 표시하시오.





UShealth <- read.table("C:\\US\_health.txt")

colnames(UShealth) = c("state", "land", "popu", "acc", "card",

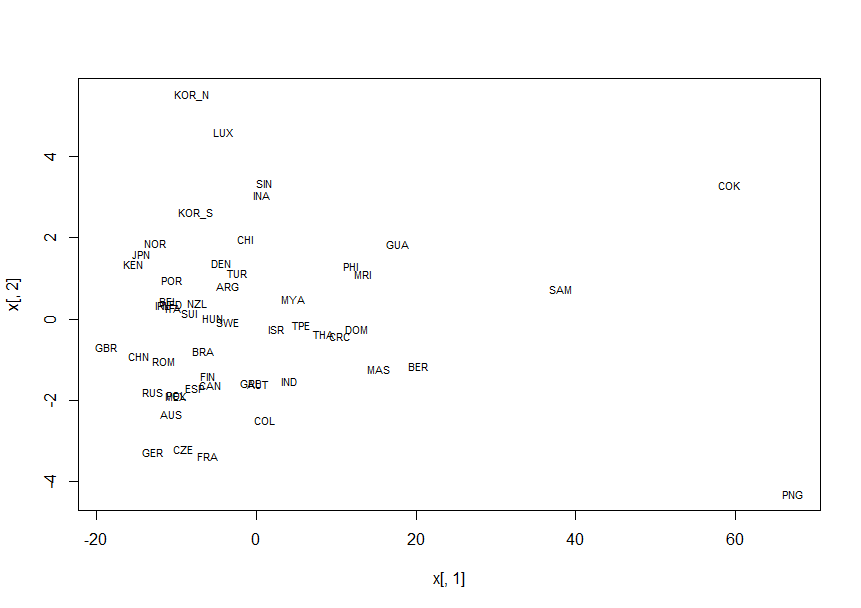
"can", "pul", "pnue", "diab", "liv",

"doc", "hosp", "r", "d")

head(UShealth)

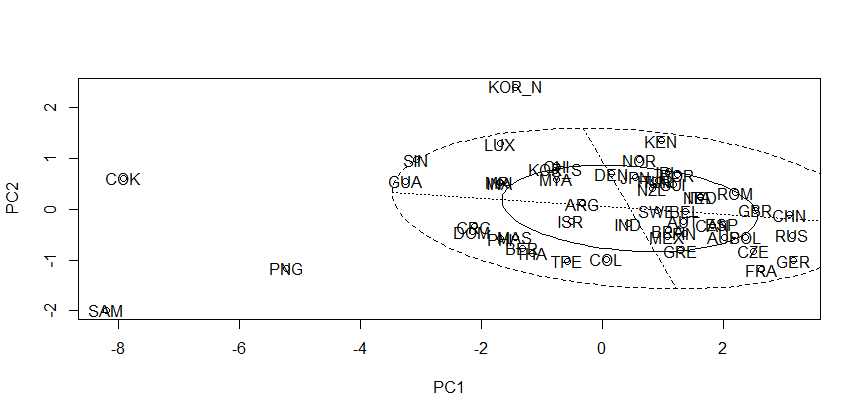
5. 2번에서 고려한 Table1\_9.txt 파일은 각 나라의 여자 육상선수들의 최고기록들을을 나타낸다. 예로서 아르헨티나(ARG)의 100m 여자 최고기록은 11.57초를 나타낸다.

a) Table1\_9.txt 파일을 이용하여 다차원척도법을 시행하고 2차원 공간에 표시하시오.

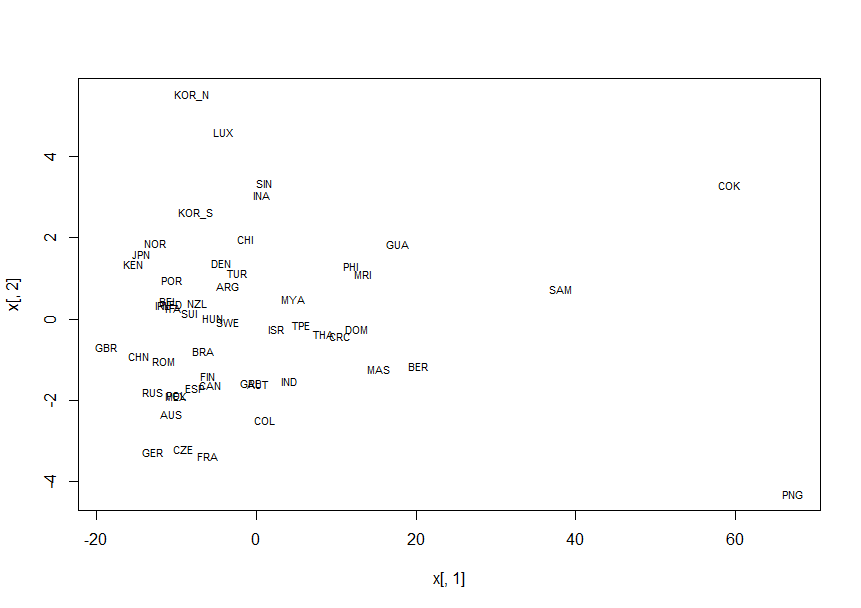


b) 2번에서 그린 첫번째 주성분점수와 두번째 주성분 점수의 산점도와 위 다차원척도법에 의해 2차원 공간에 표시한 공간을 비교하시오.

첫번째 주성분점수와 두번째 주성분 점수의 산점도



다차원척도법을 시행하여 2차원으로 표시



두 개의 2차원 공간은 서로 x축의 부호가 음수와 양수로 다르다는 것 빼고는 전체적인 모양이 비슷하다.