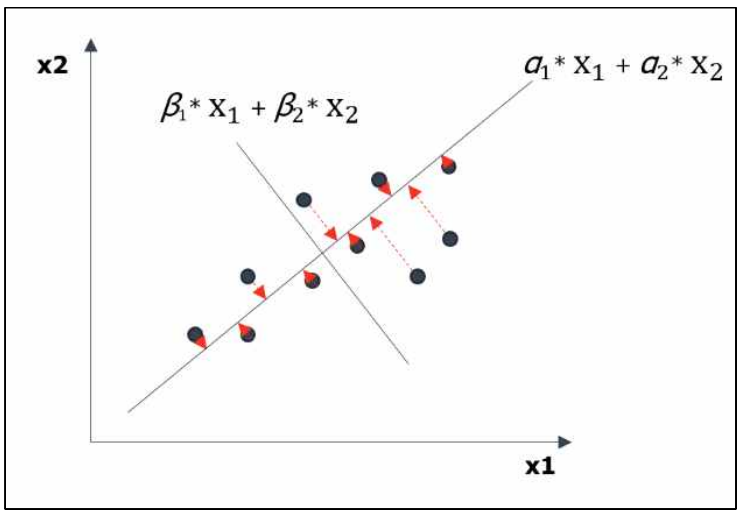
## ## 주성분 분석 PCA

- 상관관계를 갖는 많은 변수를 상관관계가 없는 소수의 변수로 변화하는 차원 축소 방법
- 장점: 데이터 복잡성 감소, 다중공선성(독립 변수의 일부가 다른 독립변수의 조합으로 설명 가능한 경우) 문제 해결이 가능함. 즉, 모델의 설명력을 높이는 동시에 복잡도를 낮추는 방법

$$PC_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{ip}x_p$$

*PCi는 i번째 주성분을, Xp는 p번째 변수, aip는 주성분 선형결합의 p번째 변수에 대한 가중치를 의미함.*

- 특징
    1. 주성분은 기존 변수의 수만큼 생성됨.
    2. 주성분은 기존 변수의 선형 결합으로 생성됨
    3. 주성분은 분산을 최대화 함
    4. 주성분끼리는 서로 독립적임(상관x)




*축 간의 평균의 교점을 원점으로 잡고, 원점을 지나는 직선에 내린 수선의 발의 총합이 최대가 되는 직선을 구함(a). 이후 해당 직선과 직교하는 두 번째 직선을 구힘(b)*

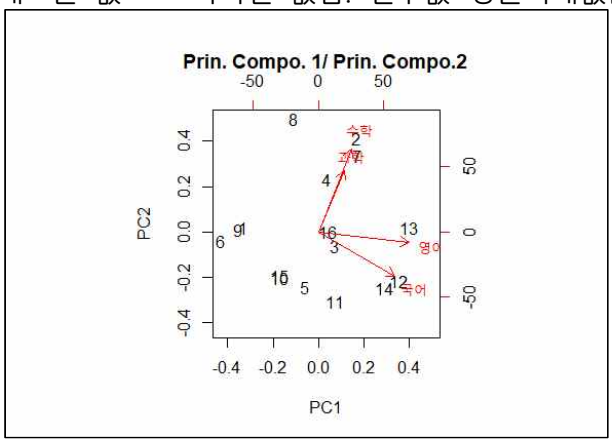
- R에서의 Cumulative proportion을 확인하고, 설명할 정도를 결정해야 함.

```
Importance of components:
                          PC1    PC2    PC3     PC4
Standard deviation     1.8971 1.2775 1.0545 0.84113
Proportion of Variance 0.4499 0.2040 0.1390 0.08844
Cumulative Proportion  0.4499 0.6539 0.7928 0.88128
```

***일부만 가져옴.*

*PC3까지 사용하면 약 79%를, PC4까지 사용하면 88% 설명 가능함*

- 성분과 변수 간의 설명 정도는 성분 적재값으로 확인해야 함.
  - 성분 적재값 : 성분과 변수 간의 상관관계를 나타냄. 크기가 클수록 성분과 변수 간의 관계가 큼을 의미함.
  - 성분 점수: 성분 적재값을 가중치로 사용하고 기존 변수들의 값을 성분값으로 표현해 변수들의 선형결합을 통해 나온 점수로, 기존 다수 변수의 관측값을 소수 성분의 새로운 값으로 축약한 값임. 변수값*성분적재값을 합산해 나타남.




*축과 평행할수록 밀접한 상관을 가짐을 의미함. 여기서 주성분 1은 영어 국어와, 주성분 2는 수학 과학 변수와 밀접한 상관을 가지고 있음.*

## ## 타당도 분석

- 요인 분석에서의 타당도는, 측정 척도가 측정하고자 하는 개념을 정확하게 나타내고 있는 지를 의미함.
- 측정 항목들이 특정 요인에 대해 0.6 이상의 요인적재값을 보이며 동시에 다른 요인에 대해선 0.3 미만의 요인 적재값을 보일 때 집중 & 판별 타당도를 갖는다고 함. (그러나, 적정 값은 표본 크기 등 바뀔 수 있음.)
    - 집중 타당도 : 같은 개념을 측정한다면 다른 측정방법을 사용했을 때에도 측정값들 간의 높은 상관관계가 존재해야 함.
    - 판별 타당도: 다른 개념을 측정한다면 다른 측정방법을 사용했을 때에도 측정값들 간의 차별성이 나타나야 함.

## ## 예제

R 내장 데이터 USArrests 를 사용하여 주성분 분석을 수행하시오.

1) 주성분의 설명 정도를 확인하고,
2) fviz_contrib 함수를 이용하여 주성분에 대한 기존 변수의 기여도를 시각화 하고,
3) 성분 적재값과 성분 점수를 확인하고
4) biplot을 생성하여 PC1, PC2 각각에 높은 상관을 가지는 변수를 확인하시오.