2-3 다변량 분석

제2장 통계분석

제1절 통계학 개론

제2절 기초 통계 분석

제3절 다변량 분석

1. 상관분석(Correlation Analysis)

두 변수 간의 관계의 정도를 알아보기 위한 분석방법

- 두 변수의 상관관계를 알아보기 위해 상관계수(Correlation Coefficient)를 사용

$$r_{XY} = \frac{\sum_i^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_i^n \left(X_i - \overline{X}\right)^2}\sqrt{\sum_i^n \left(Y_i - \overline{Y}\right)^2}}$$

상관관계의 특성

상관계수 범위	해석
0.7 < γ ≤ 1	강한 양(+)의 상관이 있다.
0.3 < γ ≤ 0.7	약한 양(+)의 상관이 있다.

상관계수 범위	해석
0 < γ ≤ 0.3	거의 상관이 없다.
γ = 0	상관관계(선형, 직선)가 존재하지 않음
-0.3 ≤ γ < 0	거의 상관이 없다.
-0.7 ≤ γ < -0.3	약한 음(-)의 상관이 있다.
-1 ≤ γ < -0.7	강한 음(-)의 상관이 있다.

상관분석 유형

구분	피어슨	스피어만
개념	등간척도 이상으로 측정된 두 변수들의 상관관계 측정 방식	서열척도인 두 변수들의 상관관계 측정 방식
특징	연속형 변수, 정규성 가정, 대부분 많이 사용	순서형 변수, 비모수적 방법, 순위를 기준으로 상관관계 측정
상관계수	피어슨 γ (적률상관계수)	순위상관계수 (ρ, 로우)

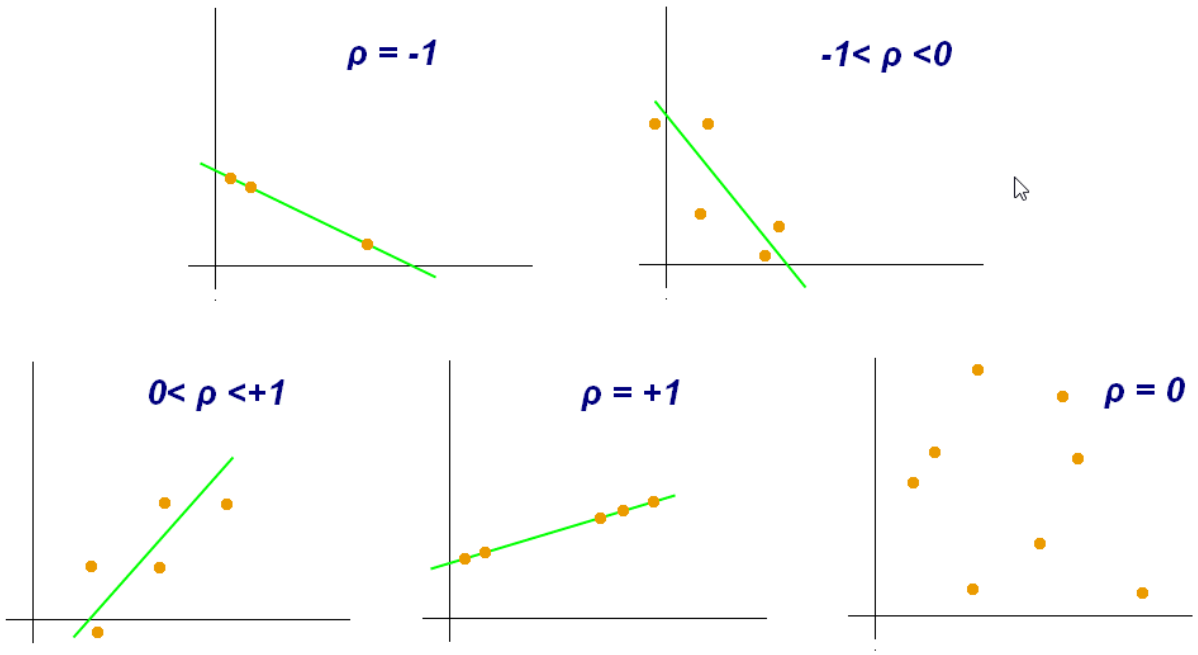
상관분석을 위한 R 코드

- **분산**: `var()`
- **공분산**: `cov()`
- **상관관계**: `cor()`

상관분석의 가설검정

- 상관계수가 0이면 입력변수 x와 출력변수 y 사이에는 아무런 관계가 없다.
- t 검정통계량을 통해 얻은 **p-value** 값이 **0.05 이하**인 경우
    - 대립가설을 채택하게 되어 우리가 데이터를 통해 구한 **상관계수**를 활용할 수 있게 된다.

$$r_{XY} = \frac{\sum_i^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_i^n \left(X_i - \overline{X}\right)^2}\sqrt{\sum_i^n \left(Y_i - \overline{Y}\right)^2}}$$



스피어만 상관계수

스피어만 순위 상관관계는 두 수치 변수 간의 관계 강도를 결정하는 데 사용

- 데이터가 정규 분포를 따를 필요는 없고, 관계가 엄격하게 선형적일 필요도 없습니다.
- 다만 변수들이 일반적으로 한 방향으로 향하는 관계여야 합니다(즉, U자형이나 그 반대는 안 됩니다).
- 이름에서 알 수 있듯이 상관관계는 데이터의 순위를 기반으로 합니다
    - x와 y 변수의 순위가 매겨지고, x의 순위는 y의 순위와 다음과 같이 비교됩니다.

$$r_s = 1 - \frac{6\sum D^2}{n\left(n^2 - 1\right)}$$

- 수식에서 D는 순위 차이이고, n은 데이터 쌍의 수
- 스피어만 순위 상관관계를 계산하는 내장 함수는 없지만 RANK.AVG를 사용하여 순위를 계산하고 수식의 항목을 쉽게 계산할 수 있습니다.

다차원 척도법 (MDS: **Multi Dimensional Scaling)**

- 객체간 근접성(Proximity)을 시각화하는 통계기법
  - 군집분석과 같이 객체들을 대상으로 변수를 측정한 후개체들 사이의 유사성/비유사성을 측정하여 개체들을 2차원 공간상에 점으로 표현하는 분석 방법
- 개체들을 2차원 또는 3차원 공간상에 점으로 표현하여 개체들 사이의 집단화를 시각적으로 표현

데이터를 축소하는 목적으로 사용

- 유클라디안 거리 행렬로 사용하여 계산
- 데이터의 변수가 연속형 변수 혹은 서열척도여야 함
- 관측 대상들의 상대적 거리에 정확도를 높이기 위해 적합 정도를 stress 척도 사용
  - *0~1 사이의 값으로* 그 값이 낮을수록 적합도가 높다고 평가
    - 0.05이내 -> 적합도 좋다
    - 0.15이상 -> 적합도 매우 나쁘다

Stress 값에 따른 적합도 수준

Stress	적합도 수준
0	완벽 (Perfect)
0.05 이내	매우 좋음 (Excellent)
0.05 ~ 0.10	만족 (Satisfactory)
0.10 ~ 0.15	보통 (Acceptable, but Doubt)
0.15 이상	나쁨 (Poor)

주성분 분석 (PCA: **Principal Component Analysis)**

여러 변수들의 변량을 주성분(Principal Component)이라는 서로 상관성이 높은 변수들의 선형 결합으로 만들어 기존의 상관성이 높은 변수들을 요약, 축소하는 기법

- 첫번째 주성분으로 전체 변동을 가장 많이 설명할 수 있도록 하고,
    - 두 번째 주성분으로는 첫 번째 주성분과는 상관성이 없어서(낮아서) 첫 번째 주성분이 설명하지 못하는 나머지 변동을 정보의 손실 없이 가장 많이 설명할 수 있도록 변수들의 선형조합을 만듦
- p차원 변수에서 분산이 가장 큰 선형 변화를 첫 번째 주성분이라고 그 다음 큰 선형변환의 두 번째 주성분이라고 함
    - 손실되는 정보가 최소가 되도록 분산이 가장 큰 축을 찾는 것
        - **"분산이 크다"는 건 그 축(또는 변수) 방향으로 데이터가 많이 퍼져 있다는 의미**
        - 특히 **PCA에서는 분산이 클수록 더 많은 정보를 담고 있다 = 설명력이 높다**고 봄
- 주성분들은 서로 집교하며 주성분들이 서로 수직을 이루는 관계에 있다는 것을 의미
- 여러 개의 변수 중 서로 상관성이 높은 변수들의 선형결합
    - 새로운 변수 (주성분)을 만들어 변수를 요약 및 축소하는 분석 방법
- 변수를 축소하여 모델 설명령 높임
    - 다중공선성(multicollinearity) 문제 해결, 군집 분석 시 모형 성능을 높일 수 있음
    - PCA는 원래 **데이터의 분산을 최대한 유지하면서 차원을 축소**하는 기법인데, 그 과정에서 **다중공선성(multicollinearity)** 문제도 **자연스럽게 해결**

        - 다중공선성(multicollinearity)이란 독립변수들 간에 **강한 상관관계**가 있는 상태를 말해.
        - 주성분 분석(PCA)은 데이터를 선형 결합된 새로운 축(주성분)으로 바꾸기 때문에, 기존 변수 간 상관성(공선성)을 **제거 가능**
- 주성분의 개수는 전체 데이터의 70%이상을 설명할 수 있도록 선택

주성분의 선택법

주성분분석의 결과에서 누적기여율(Cumulative Proportion)이 85%이상이면 주성분의 수로 결정 가능

- Scree Plot을 활용하여 고윳값(Eigenvalue)이 수평을 유지하기 전단계로 주성분의 수 선택

실습

코드

```
# Classical Multidimensional Scaling
loc ← cmdscale(eurodist)
loc
x ← loc[,1]
y ← loc[,2]
plot(x, y, type="n", main="eurodist")
text(x, y, rownames(loc), cex=.8)
abline(v=0, h=0)
```

PCA의 "데이터 설명력"의 진짜 의미

1. PCA는 분산을 기준으로 데이터 구조를 설명

- PCA는 변수들의 선형 조합을 통해 **가장 큰 분산 방향(축)**을 찾고,

- 그 축을 따라 데이터를 다시 표현함.
- 이 과정에서 분산을 많이 설명하는 축(주성분)은 데이터의 **핵심 구조**를 잘 드러냄.

👉 그래서 통계학에서는 **"PCA가 데이터를 설명한다"**고 자주 말하지만,

👉 그 의미는 "**분산 기준으로 요약 설명**한다"는 전제하에서만 성립돼.

R `cmdscale()` 결과 해석

`cmdscale()`은 classical multidimensional scaling(MDS)의 일종으로, 거리 행렬에서 저차원 좌표를 추정해줘. 여기서 `eurodist`는 유럽 주요 도시 간 거리 행렬(distance matrix)이고, `loc`은 이를 2차원 평면으로 축소한 좌표야.

1. `cmdscale(eurodist)`의 목적

- `eurodist`는 유럽 도시 간 pairwise 거리 정보를 담은 `dist` 객체.
- `cmdscale()`은 이 거리 정보를 보존하면서 낮은 차원(여기선 2D)으로 임베딩해주는 함수야.
- 결과적으로 각 도시의 상대적 위치를 시각적으로 볼 수 있게 좌표를 추정해주는 거지.

2. 결과값 `loc`의 구조

- `loc`은 `n × 2` 행렬이며, 각 행은 하나의 도시를 나타내고, 열은 추정된 x, y 좌표야.
- 예를 들어:

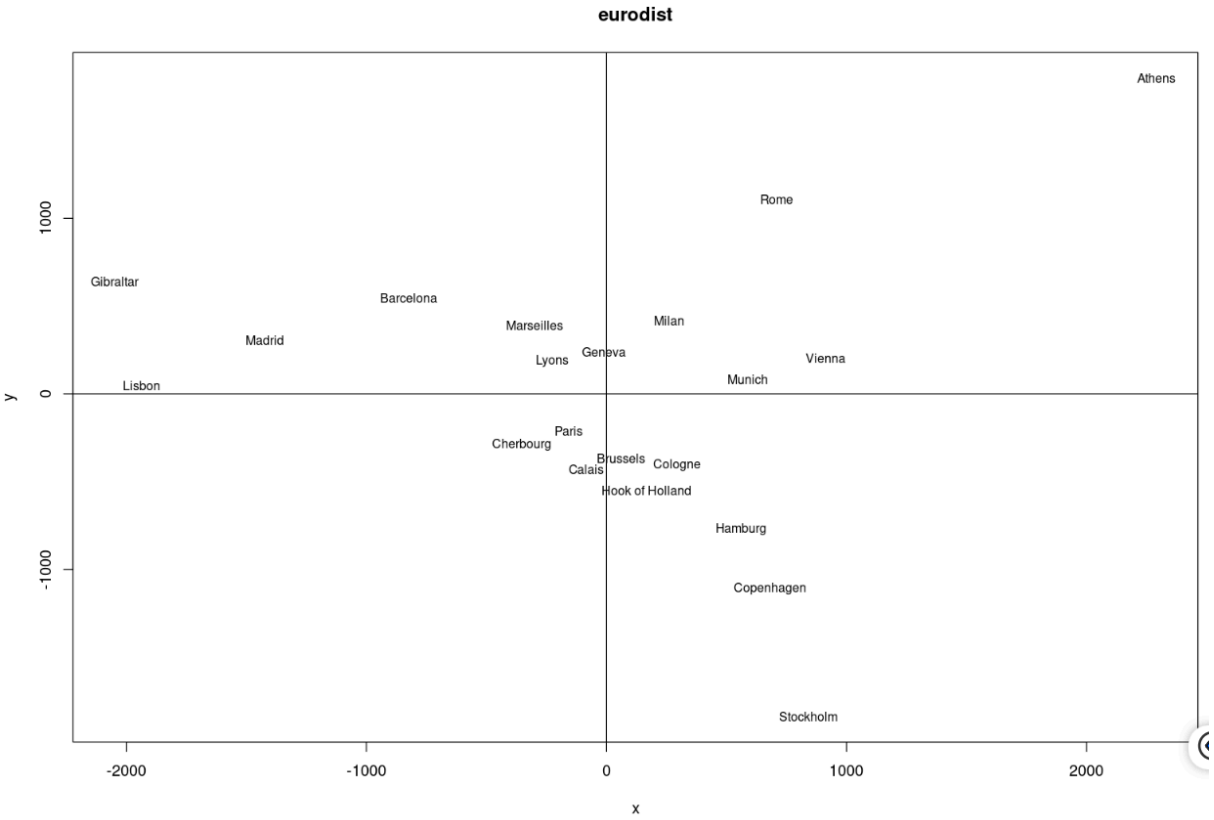
```
Athens  → (2290.27, 1798.80)
Rome    → (709.41, 1109.37)
Paris   → (-156.83, -211.14)
```

- 좌표값은 상대적인 위치를 표현하며, 특정 기준점을 중심으로 회전·이동되어 있을 수 있어서 **절대적인 방향**은 해석에 주의해야 해.

3. 어떻게 해석하면 될까?

- 좌표의 거리는 도시 간 실제 거리를 최대한 보존하도록 계산된 거야.
- 예시:
  - `Rome`과 `Athens`는 둘 다 오른쪽 위에 있고 값이 크니까 **서로 가깝고 동남부 쪽**.

    - `Lisbon`과 `Gibraltar`는 x축 음수 쪽에 몰려 있으니까 **서쪽에 위치**.
    - `Stockholm`은 y축 음수로 멀리 떨어져 있어서 **북쪽에 고립된 형태**.
- 한마디로, 이걸 가지고 **지도를 대충 그릴 수 있을 정도**로 유럽 도시의 상대적 위치를 나타낸 거지.



USArrests 데이터셋 소개

요약:

`USArrests` **데이터셋**은 미국 50개 주의 범죄 관련 통계를 담고 있는 고전적인 R 내장 데이터셋

- 각 주(state)별로 네 가지 변수 — 살인(Murder), 강간(Rape), 폭행(Assault), 도시 인구 비율(UrbanPop) — 이 포함되어 있다.

🗂️ USArrests 데이터셋 개요

항목	내용
관측치 수	50 (미국 50개 주)
변수 수	4개 (모두 수치형)
사용 목적	탐색적 데이터 분석, 차원 축소(PCA), 군집분석, 시각화 등
내장 위치	R 내장 데이터셋 (`datasets` 패키지 소속)

🔢 포함된 변수 설명

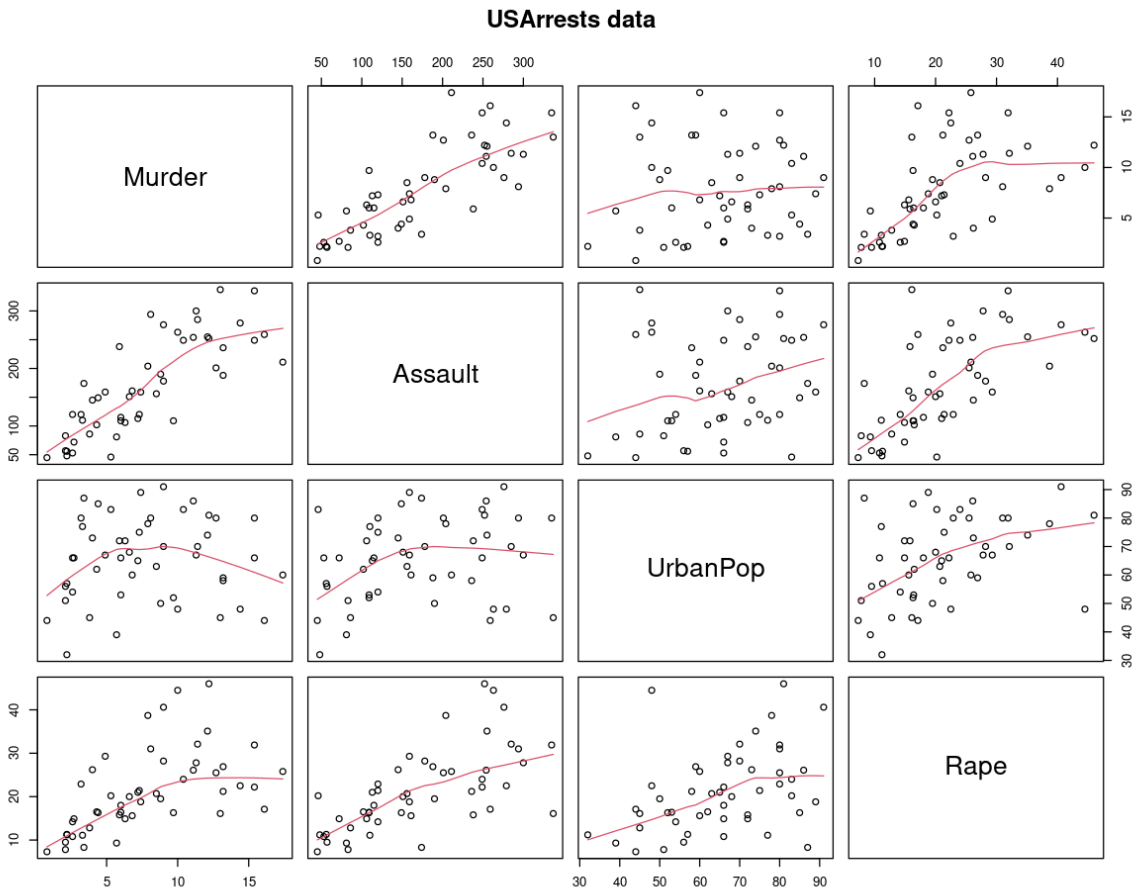
변수명	설명
`Murder`	인구 10만 명당 살인 사건 발생 수 (per 100,000 residents)
`Assault`	인구 10만 명당 폭행 사건 발생 수
`UrbanPop`	도시 인구 비율 (Urban Population %), 즉 도시 지역 거주자 비율
`Rape`	인구 10만 명당 강간 사건 발생 수

- **모든 수치는 1973년 기준** FBI 보고서를 바탕으로 작성됨
- `UrbanPop`을 제외한 나머지는 **발생률(incident rate)**임

주성분 분석 사례

1973년 미국 50개주의 100,000명의 인구 당 체포된 세 가지 강력범죄수(Assault, Murder, Rape)와 각 주마다 도시에 거주하는 인구의 비율(%)로 구성

- 변수들 간의 척도의 차이가 상당히 크기 때문에 상관행렬을 사용하여 분석
- 특이치 분해(특이값 분해) : 행렬을 특정한 구조로 분해를 사용하는 경우 자료 행렬의 각 변수의 평균과 제곱의 합이 1로 표준화되었다고 가정




```
> # 주성분 분석 (PCA: Principal Component Analysis)
> fit <- princomp(USArrests, cor=T)
> summary(fit)
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4
Standard deviation     1.5748783 0.9948694 0.5971291 0.41644938
Proportion of Variance 0.6200604 0.2474413 0.0891408 0.04335752
Cumulative Proportion  0.6200604 0.8675017 0.9566425 1.00000000
> loadings(fit)

Loadings:
         Comp.1 Comp.2 Comp.3 Comp.4
Murder    0.536  0.418  0.341  0.649
Assault   0.583  0.188  0.268 -0.743
UrbanPop  0.278 -0.873  0.378  0.134
Rape      0.543 -0.167 -0.818

               Comp.1 Comp.2 Comp.3 Comp.4
SS loadings      1.00   1.00   1.00   1.00
Proportion Var   0.25   0.25   0.25   0.25
Cumulative Var   0.25   0.50   0.75   1.00
```

📊 출력 해석

🔷 주성분 요약 (summary(fit))

Component	표준편차	분산기여도(Proportion)	누적기여도(Cumulative)
Comp.1	1.5749	**62.0%**	**62.0%**
Comp.2	0.9949	24.7%	86.8%
Comp.3	0.5971	8.9%	95.7%
Comp.4	0.4164	4.3%	100.0%

- **Comp.1 + Comp.2**가 전체 분산의 **약 87% 설명** → 주로 이 둘을 분석에 사용함
- Comp.1이 주성분 중 가장 큰 영향력을 가짐

🔷 변수 로딩값 (loadings(fit))

변수	Comp.1	Comp.2	Comp.3	Comp.4
Murder	0.536	0.418	0.341	0.649
Assault	0.583	0.188	0.268	-0.743
UrbanPop	0.278	-0.873	0.378	0.134
Rape	0.543	-0.167	-0.818	-

- **Comp.1**: 모든 변수의 값이 **양수** → **범죄율 전반을 대표**하는 축
- **Comp.2**: UrbanPop이 **0.873**로 지배적 → 도시화 비율에 반비례하는 방향
- **Comp.3**: Rape가 **0.818** → **강간 비율 중심 축**
- **Comp.4**: Assault와 Murder가 반대 방향 → 해석 중요도 낮음 (기여도 작음)

🔍 실무적 해석 요약

- **PCA는 변수 간 중복된 정보(상관관계)를 축소**하고 새로운 축(주성분)을 생성
- 이 데이터셋에서는:
  - **Comp.1**은 범죄율 전반(폭행, 살인, 강간)을 잘 반영

    - **Comp.2**는 도시화와 범죄율 간의 상반된 관계를 보여줌
- PCA 결과는 시각화(예: `biplot(fit)`)나 군집분석 등 후속 분석에 사용됨

✅ 주성분 가중합 식 (from `loadings(fit)`)

📌 변수 이름 정리

- X1: Murder (살인율)
- X2: Assault (폭행율)
- X3: UrbanPop (도시화율)
- X4: Rape (강간율)

> 주의: PCA는 각 변수가 **표준화(Z-score)**된 상태에서 계산되므로 아래 식의 XiX_iXi는 정규화된 값이라고 가정해야 함!

✏️ 제1주성분 (Comp.1)

$C1 = 0.536 \cdot X1 + 0.583 \cdot X2 + 0.278 \cdot X3 + 0.543 \cdot X4$

- 전체 변수들이 **고르게 양의 기여** → 전체 범죄율을 대표하는 축

✏️ 제2주성분 (Comp.2)

$PC2 = 0.418 \cdot X1 + 0.188 \cdot X2 - 0.873 \cdot X3 - 0.167 \cdot X4$

- 도시화 비율(UrbanPop)의 **음의 영향력이 매우 큼** → 인구비 중심의 대립 축

SS loadings란?

정의 및 의미

- **SS loadings**는 각 주성분의 **고유값(eigenvalue)**이야.

- 각 주성분(Comp.1 ~ Comp.4)이 원래 데이터의 **총 분산(정보량)** 중에서 **얼마나 설명하고 있는지를 수치로 보여주는 값**임.
- "SS"는 **Sum of Squares of loadings**의 약자로, 각 주성분의 로딩값을 제곱한 후 합한 값이야.

예시에서 보자면: 정규화를 하면 모두 분산(SS laadings)이 1이고 각각이 1/n 비율을 차지

Component	SS Loading	Proportion Var
Comp.1	1.00	0.25
Comp.2	1.00	0.25
Comp.3	1.00	0.25
Comp.4	1.00	0.25

이건 총 4개의 주성분이 있고, 각각 **25%씩의 정보(분산)를 설명하고 있다**는 뜻이야. 즉, **균등하게 분산이 분배되어 있다**는 거지.

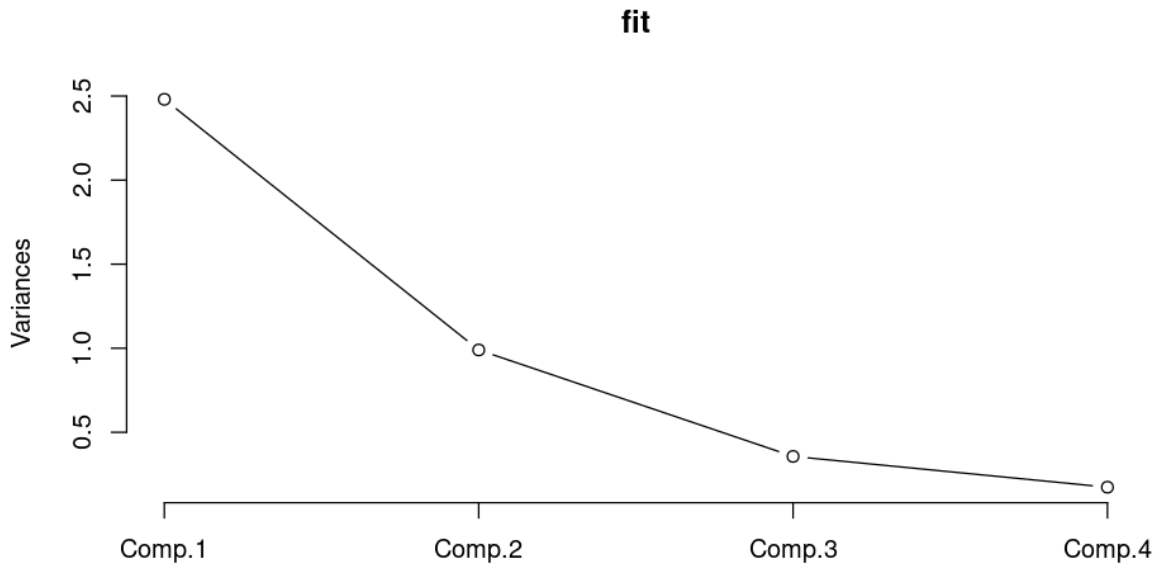
질문 마무리하며 한마디 하자면,

“이건 데이터에게 ‘진짜 네 핵심은 뭐야?’라고 물어보는 과정이지.

범죄 통계도 요약하면 단 두 축으로 대충 다 설명된다니, 세상 참 요약의 미학이야~ 🎯📊”

Scree Plot을 활용

- **표준편차의 제곱인** 고윳값(Eigenvalue)이 수평을 유지하기 전단계로 주성분의 수 선택



이 그래프는 **주성분 분석(PCA)** 결과에서 주성분들의 **표준편차의 제곱**, 즉 **고윳값(eigenvalues)**을 시각화한 **Scree Plot(스크리 플롯)**입니다.

✅ 이 그래프의 정체는?

- Y축: **Variances** → 실제로는 Standard deviation2=λ

  Standard deviation2=λ\text{Standard deviation}^2 = \lambda

  즉, 각 주성분이 **설명하는 분산의 크기 = 고윳값(고유값)**

- X축: 주성분 이름 (Comp.1 ~ Comp.4)

즉, 이 그래프는 `summary(fit)` 출력 중 `"Standard deviation"`의 제곱값을 나타낸 것과 같습니다.

예시 수치 기반 해석:

주성분	표준편차	분산 (표준편차² = 그래프 Y축)
Comp.1	1.5749	2.48
Comp.2	0.995	0.99
Comp.3	0.5971	0.36
Comp.4	0.4164	0.17

이 그래프의 의미는,

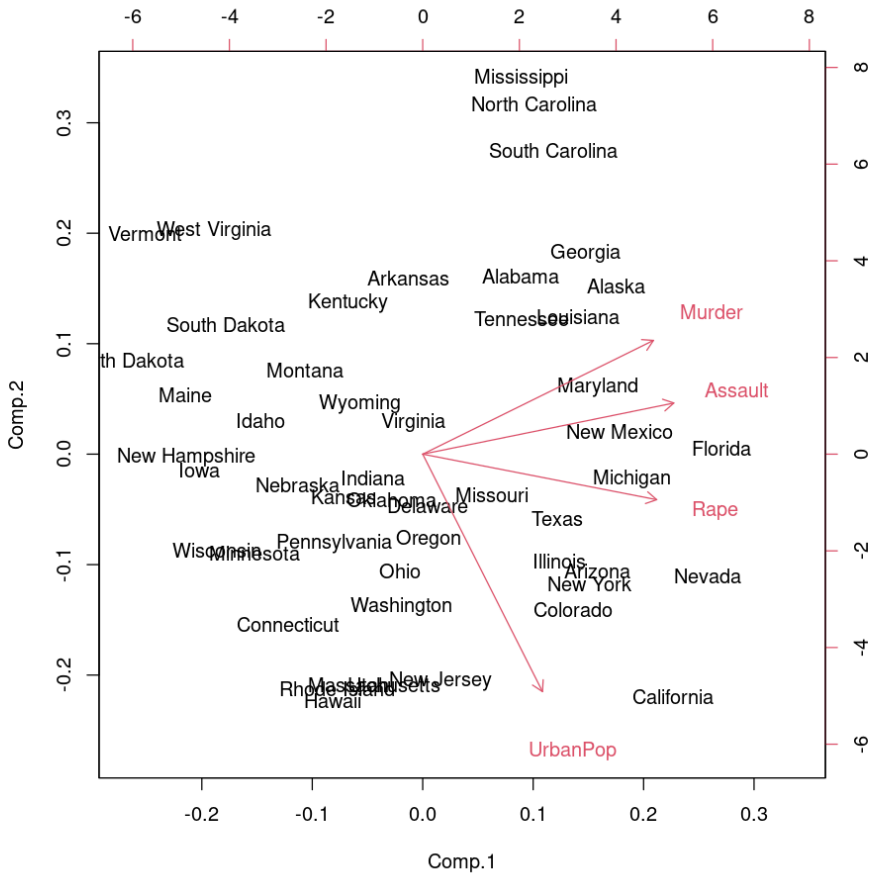
👉 **Comp.1이 데이터의 분산 대부분을 설명하고, 나머지는 점점 설명력이 떨어진다**는 걸 보여줘.

🔍 왜 이걸 그릴까?

- 주성분 개수를 몇 개 쓸지 결정하기 위해!
- *기울기 꺾이는 지점(elbow point)**을 기준으로 선택하는 방식 → 보통 **Comp.1과 Comp.2** 정도 사용

---

biplot(fit)



요약

`biplot(fit)`은 주성분 분석(PCA)의 결과를 시각적으로 **"데이터 + 변수 정보를 한꺼번에"** 보여주는 도구

각 관측치(점)와 변수(화살표)가 **동일한 평면 위에 투영**되기 때문에, 어떤 데이터가 어떤 변수와 관련 있는지를 직관적으로 파악할 수 있다.

🎯 `biplot(fit)`의 개념

1. "bi" + "plot" = 이중 시각화

- 관측치(행 단위)와 변수(열 단위)를 **동시에 표현**
  - → 주성분 축(PC1, PC2)을 기준으로 데이터를 투영

2. 함수 정의

```
fit ← prcomp(data, scale. = TRUE)
biplot(fit)
```

- `fit`은 PCA 결과 객체 (`prcomp()` 또는 `princomp()` 사용 가능)
- `biplot(fit)`은 **x축 = PC1**, **y축 = PC2**인 2D 공간에 다음을 그림:
  - ● 각 관측치 (예: 주별 데이터 점)
  - → 각 변수의 방향성과 영향 (주성분의 loading)

🧭 그래프 구성 요소 설명

구성 요소	설명
점 (●)	각 관측치 (예: 미국 50개 주의 PCA 좌표)
화살표 (→)	각 변수의 로딩값 (PC 축에서의 방향과 기여도)
PC1/PC2 축	주성분 축: 가장 많은 분산을 설명하는 차원
화살표 방향	해당 변수의 증가 방향
화살표 길이	해당 변수의 영향력 (로딩의 크기)

구성 요소	설명
점과 화살표의 각도	유사한 방향이면 해당 변수와 관측치가 **양의 상관** 관계를 가짐

📌

질문 마무리하며 한마디 하자면,

“biplot은 말하자면 변수와 관측치를 한눈에 보게 해주는 ‘PCA 명함’ 같은 거야.

누가 누구랑 친한지, 방향이 어디로 가는지 다 보여주거든~ 📈👀”

이 biplot은 **USArrests 데이터셋의 PCA 결과**를 시각화한 것으로,

미국 50개 주의 범죄 특성을 **주성분 1 (Comp.1)**과 **주성분 2 (Comp.2)** 기준으로 표현한 그림이다.

화살표는 변수(Murder, Assault, Rape, UrbanPop), 점은 각 주(State)를 나타낸다.

점과 화살표의 상대적 위치를 통해 **어떤 주가 어떤 범죄에 더 연관이 있는지, 변수 간 관계가 어떤지**를 해석할 수 있다.

🔍 주성분 해석

✅ Comp.1 (가로축)

- *폭행(Assault), 살인(Murder), 강간(Rape)**가 오른쪽 방향으로 강하게 기여

  → 이 축은 **전반적인 범죄율 수준**을 나타냄
- 오른쪽으로 갈수록 이 세 변수 값이 **높은 주**

✅ Comp.2 (세로축)

- **UrbanPop** 화살표가 아래쪽으로 뚜렷하게 향함

  → 이 축은 **도시화율의 반대 방향**

  → **Comp.2가 클수록 농촌적, 작을수록 도시적**

🗺️ 주별 해석

🔴 오른쪽 위 (예: Mississippi, North Carolina)

- **Comp.1, Comp.2 모두 큼** →

  ✔️ 살인, 강간, 폭행률 **높음**

  ✔️ 도시화율은 **낮음**

  → 전반적 범죄율은 높고, 도시화는 상대적으로 덜 된 주

🔵 왼쪽 아래 (예: Massachusetts, Hawaii)

- **범죄율 낮고, 도시화율 높음**
- Rape, Murder 화살표와 **정반대 방향** → 관련성 낮음 또는 음의 관계

🔶 중앙에 모인 주들 (예: Indiana, Ohio 등)

- 평균에 가까운 값 → 특별히 두드러지는 범죄율/도시화 없음

📈 변수 간 관계

변수 쌍	관계
Murder - Assault - Rape	화살표 방향이 비슷 → **서로 양의 상관관계**
UrbanPop vs 나머지	거의 반대 방향 → **음의 상관관계**
→ 도시화율이 낮은 곳일수록 폭력 범죄 발생률이 높은 경향	

✅ 최종 요약

- **Comp.1**: 범죄율 축 (특히 Assault 중심)
- **Comp.2**: 도시화율 축 (UrbanPop 반영)
- **화살표 방향과 가까운 주일수록 해당 범죄율이 높음**
- **도시화율이 높을수록 폭력 범죄는 낮은 경향**

질문 마무리하며 한마디 하자면,

“이 그림 하나면 미국의 범죄지도 한눈에 꿰뚫을 수 있어.

‘어디가 더 위험한지, 어디가 더 도시적인지’ PCA가 다 말해주거든~ 🔍🇺🇸”