MDA?

- 좁은 의미의 다변량 분석
  - 변수가 3개 이상인 데이터 분석 방법
  - 인과 관계 분석은 제외, no Y(종속변수)

$$
\begin{array}{cc}
 & \leftarrow \text{측정변수} \rightarrow \\
\uparrow \\ \text{개체} \\ \downarrow &
\begin{bmatrix}
x_{11} & x_{12} & \cdots & x_{1p} \\
x_{21} & x_{22} & \cdots & x_{2p} \\
 & \cdots\cdots\cdots\cdots & & \\
x_{n1} & x_{n2} & \cdots & x_{np}
\end{bmatrix}
\end{array}
$$

개체 분류
- 군집 분석
- 판별 분석

데이터 축약
- 주성분 분석
- 요인분석

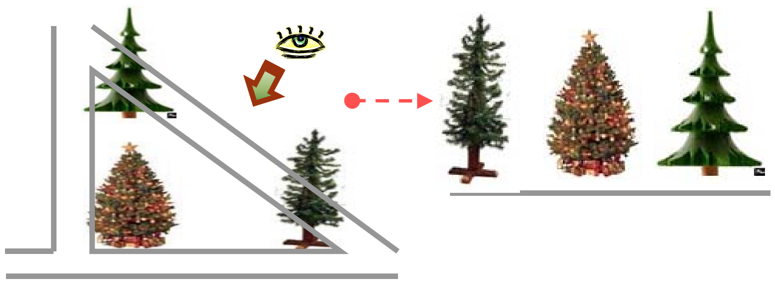
- 데이터 차원 축약: 변수들의 상관관계 활용
  - 주성분분석 (principal component analysis)
  - 요인분석 (factor analysis)
- 개체 분류: 개체 유사성(거리) 이용
  - 군집분석 (clustering analysis)
  - 판별분석 (discriminant analysis)
- Other MDA
  - 정준상관분석 (canonical correlation): 변수 그룹간 상관분석
  - 대응분석 (correspondence analysis): 개체 분류 범주 분류

- Multivariate normal using R
  - Definitely, Sometimes, Rare, Possible, Not possible

	주성분 분석	요인 분석	판별 분석	군집 분석	정준 상관 분석
변수들 관계 탐색	S	D	N	N	S
자료 탐색	D	S	N	S	N
새 변수 만들기	Yes	Yes	No	No	Yes
개체 분류	No	No	Yes	Yes	No
그룹간 평균 비교	P	P	R	R	No
변수 그룹	P	P	N	N	D
차원(변수) 줄이기	D	P	N	N	P

PCA Concept

- 주성분 개념: 차원의 축약
  - 어느 곳에서 바라보면 희생되는 정보가 가장 적은가?



  - 주성분 변수
    - 원변수의 선형 결합
    - 제일 주성분이 원변수의 변동을 가장 많이 설명: 원 변수 공분산 행렬 (상관계수 행렬)이 시작 단계

- Typical example
  - 기성복
  - Textbook 페이지 71
    - 주성분의 원변수 설명력=고유치
    - 선형계수: 고유벡터

$$S = \hat{\Sigma} = \begin{pmatrix} 518 & 4.7 \\ 4.7 & 1.22 \end{pmatrix}$$

$$\lambda_1 = 518.69, \lambda_1 = 1.18$$

$$Y_1 = l_{11} \times Weight + l_{12} \times IQ$$
$$Y_2 = l_{21} \times Weight + l_{22} \times IQ$$

- 예제2: 동일 가격 세가지 단말기에 대한 평가표의 예

단말기 명	디자인	기능성	디자인 + 기능성 (합성변량1)	디자인 – 기능성 (합성변량2)
A	40	80	120	-40
B	80	40	120	40
C	65	65	130	0

  - 합성변량 1 : 종합성, 특징이 겹치거나 단말기간 특징 차이 작음
  - 합성변량 2 : 개성 (?)A는 기능 중시, B는 디자인 중시
  - 개성 변량 분산은 1,600으로 종합 변량의 분산 33.3보다 48배 큼

- 주성분 분석은...

이처럼 "분산을 최대"로 하는 방향(변량)을 찾아 그로부터 데이터를 축약하여 데이터의 개개 정보를 보다 보기 쉽게 표현해주는 분석

- 활용(페이지 72): 주성분 분석은 중간 단계이다. (데이터 축약을 통한 개체 분류나 변수 구조 탐색 도구)
  - 데이터 스크린: 이상치, anomaly, 개체 순위, 이름부여
  - 군집 이름 부여: 군집분석 후 군집에 적절한 이름 부여
  - 판별분석
  - 회귀분석: 다중공선성 문제 해결

http://wolfpack.hnu.ac.kr

주성분 (변수) 개념 (페이지 73)

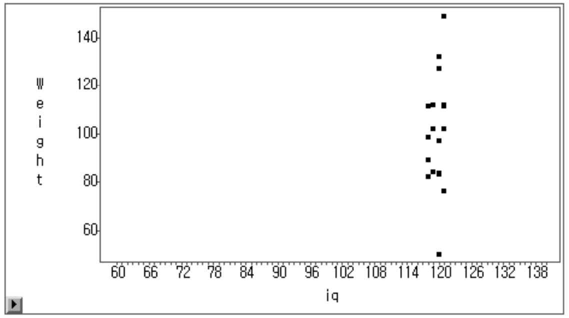
▪주성분 (변수)?

- 주성분은 원변수의 선형결합이다. (합성변량) Y=LX
- 산형계수 행렬 L: 부하행렬(loading matrix)

$$\underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} = \begin{pmatrix} l_{11}\ l_{12} \cdots l_{1p} \\ l_{21}\ l_{12} \cdots l_{2p} \\ \vdots \\ l_{p1}\ l_{p2} \cdots l_{pp} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} = L\underline{x}$$

▪원칙

- 원변수의 선형결합
- 주성분 변수간에는 공통 정보 없음, 상관계수 0
- 제일 주성분의 원 변수 변동 설명력 가장 큼
- 원변수의 개수만큼 주성분 변수가 존재: : 원변수 변동을 설명하는 비율 누적 80%인 주성분만 이용



$$y_1 = 0.999(Weight - \bar{W}) + 0.009(IQ - \bar{I})$$
$$y_2 = -0.009(Weight - \bar{W}) + 0.999(IQ - \bar{I})$$

▪개념

- p개 원 변수의 선형 결합인 주성분 변수를 이용하여 원 변수의 공분산 구조를 설명하는 방법
- 공분산 구조를 설명한다는 것은 원 변수의 변동 합과 주성분 변수의 변동 합은 동일하다는 것을 의미한다.

▪주성분 기여율

- 주성분 변수의 원변수 총 변동울 설명하는 정도
- 원변수의 종변동은 $\sum_i V(x_i) = \sum_i \hat{\sigma}_{ii} = \sum_i s_{ii}$
- 각 주성분이 원변수의 변동을 설명하는 부분은 겹치지 않는다. 그러므로 각 주성분의 원변수 변동 설명의 합은 원변수 변동과 동일하다. $V(y_i) = \lambda_i$

Covariance Matrix

	Weight	iq
Weight	518.6520468	4.7309942
iq	4.7309942	1.2280702

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	518.695300	517.510484	0.9977	0.9977
2	1.184817		0.0023	1.0000

Eigenvectors

	Prin1	Prin2
Weight	0.999958	-.009142
iq	0.009142	0.999958

http://wolfpack.hnu.ac.kr

Chapter 4. Principal Component Analysis

예제 데이터

- APPLICANT.TXT
  - 지원자 48명 (n=48)
  - 15개 항목 능력 평가(10점 만점 리커트 척도) (p=15)
  - 측정변수 속성 및 단위 모두 동일: 공분산 행렬 이용
- 우수 지원자 5명 선발
  - 지원자 업무 능력을 평가할 수 있는 단일 지표 만들기: 주성분 변수가 능력 평가 지표가 된다. 주성분 변수는 서로 독립적 구성 개념(construct)
  - 업무능력에 의한 지원자 시각적 표현: 군집분석

```
> app=read.table("Applicant.txt",header=T)
> app
  ID X1.FL. X2.APP. X3.AA. X4.LA. X5.SC. X6.LC.
1  1      6       7      2      5      8      7
2  2      9      10      5      8     10      9
```

Untitled - R Editor

```
app=read.table("Applicant.txt",header=T)
app
```

  - 변수명에서 (은.으로 자동 변환
  - R editor: 프로그램 일괄 실행 가능, CTRL+A ➔ CTRL+R

- BIG8.TXT
  - 미국 BIG8 리그 소속 대학 (n=8)
  - 수비 능력(Rushing Defense, Passing Defense) 공격 능력(Rushing Offense, Passing Offense) 획득점수(offense score) 잃은 점수(defense score) 게임, 이긴 회수
  - 측정변수의 속성, 단위 다름: 상관계수 행렬 이용
- 빅 8리그 경기 결과(승률)와 비교
  - Football 능력을 평가하는 지표 만들기, 그 지표를 이용하여 Football 경기 능력 측정
  - Football 경기 능력에 의한 팀별 시각적 표현: 군집분석

```
> big8=read.table("Big8.txt",header=T)
> big8
      SCHOOL GAMES RO_YDS RD_YDS PO_YDS
1   COLORADO    11  291.5  114.2  203.8
2 IOWA_STATE    11  178.0  272.8  137.1
3     KANSAS    11  247.1  171.2  140.9
```

Untitled - R Editor

```
big8=read.table("Big8.txt",header=T)
big8
```

http://wolfpack.hnu.ac.kr

Chapter 4. Principal Component Analysis

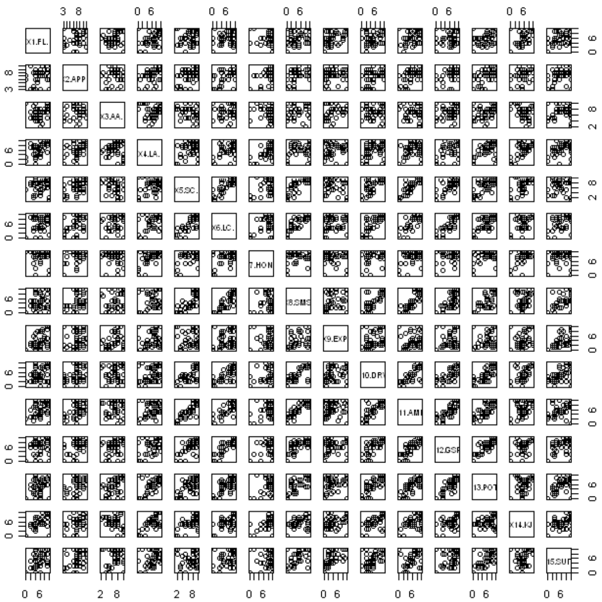
공분산 행렬 및 상관계수 행렬

▪ APPLICANT.TXT

D:₩Temp₩Applicant.R - R Editor

```
app=read.table("Applicant.txt",header=T)
app_s=app[,2:16]
round(cov(app_s),2)
plot(app_s)
```

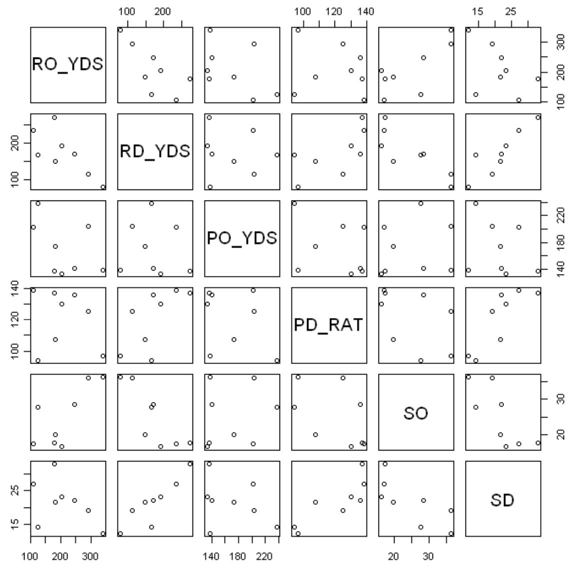
```
          X1.FL. X2.APP. X3.AA. X4.LA. X5.SC.
X1.FL.      7.15    1.26   0.23   2.30   0.60
X2.APP.     1.26    3.87   0.48   2.09   2.05
X3.AA.      0.23    0.48   3.95   0.01   0.01
X4.LA.      2.30    2.09   0.01   7.87   2.05
X5.SC.      0.60    2.05   0.01   2.05   5.85
```



▪ BIG8.TXT

```
Big8=read.table("Big8.txt",header=T)
Winpct=Big8$WINS/Big8$GAMES
Big8=cbind(Big8,Winpct)
Big8s=cbind(Big8[,3:6],Big8[,9:10])
round(cor(Big8s),2)
plot(Big8s)
```

```
> Big8
         SCHOOL GAMES RO_   TOM WINS    Winpct
1      COLORADO    11  2  0.55 10.0 0.9090909
2    IOWA_STATE    11  1 -0.64  0.0 0.0000000
3        KANSAS    11  2  0.73  6.0 0.5454545
4  KANSAS_STATE    11
```



주성분 (계수) 구하기 (페이지 75)

$$\underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} = \begin{pmatrix} l_{11}\ l_{12} \cdots l_{1p} \\ l_{21}\ l_{12} \cdots l_{2p} \\ \vdots \\ l_{p1}\ l_{p2} \cdots l_{pp} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} = L\underline{x}$$

▪주성분 구하기

- 제일 주성분 (first principal component)
  - $\underline{a}_1'\underline{a}_1 = 1$ 을 만족하는 벡터 중 $V(\underline{a}'_1(\underline{x}-\underline{\mu}))$ 을 최대화 하는 벡터 $\underline{a}_1$를 선형계수로 하여 구해진 합성변수. $\underline{y}_1 = \underline{a}_1'(\underline{x}-\underline{\mu})$
  - 공분산 행렬(S)로부터 얻어진 고유치 λ1 에 대응하는 고유벡터 e1 중 $\underline{e}_1'\underline{e}_1 = 1$을 만족하는 고유벡터 $\underline{l}_1' = \underline{e}_1'$
- 제이 주성분 (second PC)
  - $\underline{a}_1'\underline{a}_2 = 0, \underline{a}_2'\underline{a}_2 = 1$ 을 만족하는 벡터 중 $V(\underline{a}'_2(\underline{x}-\underline{\mu}))$을 최대화 하는 벡터 $\underline{a}_2$를 선형계수로 하여 계산한 합성변수.
  - 공분산 행렬(S)로부터 얻어진 고유치 λ2에 대응하는 고유벡터 e2 중 $\underline{e}_1'\underline{e}_2 = 0, \underline{e}_2'\underline{e}_2 = 1$ 을 만족하는 고유벡터
- 이와 같은 방법으로 순차적으로 구한다. $\underline{l}_2' = \underline{e}_2'$

- 원변수 벡터 공분산 행렬 (covariance matrix, Σ, S)

$$\underline{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \ldots\sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \ldots\sigma_{2p} \\ \ldots & & \\ \sigma_{p1} & \sigma_{p2} & \ldots\sigma_{pp} \end{bmatrix}$$

▪공분산 행렬(Σ)의 고유치

- 계산 방법: $|\Sigma_{p\times p} - \lambda I_p| = \underline{0}$ 을 만족하는 λ들을 고유치라 한다. $\lambda_1 \geq \lambda_2 . \geq .. \geq \lambda_p$
- 고유치는 행렬의 차수만큼 존재한다.
- 고유치 의미: 주성분 변수의 분산 (슬라이드 9에서 타원의 폭과 높이에 해당)

▪공분산 행렬(Σ)의 고유벡터

- 계산방법: $\Sigma\underline{e}_i = \lambda_i\underline{e}_i$ 을 만족하는 벡터 e를 고유벡터라 한다.
- 고유벡터는 무수히 많이 존재
- 고유벡터는 서로 orthogonal 하다. $\underline{e}_i'\underline{e}_j = 0 \;\; for\ i \neq j$
- 고유벡터를 주성분 계산 선형계수로 사용한다.

$$\underline{l}_j' = \underline{e}_j'$$

주성분 (계수) 구하기 (2)

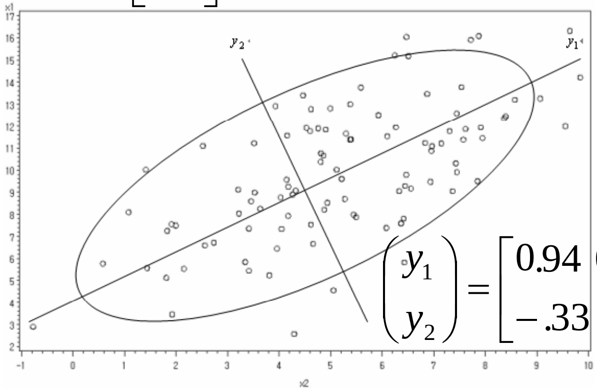
- 상관계수 행렬(correlation matrix, R) 사용 (페이지 80)
  - 원변수의 측정 단위가 상이한 경우: 주성분 계산 시 단위 크기가 큰 원변수의 영향(분산이 크므로)이 크다.
  - 문제 해결을 위하여 상관계수 행렬 사용하여 고유치, 고유벡터를 구한다.
  - 주성분 구하는 절차는 공분산 행렬과 동일하다.

- 제 k 주성분 기여율 (페이지 79)

$$\frac{V(y_1)}{S_{11}+S_{22}+\ldots+S_{pp}} = \frac{\lambda_k}{\sum_{i=1}^{p}\lambda_i}(S\text{사용})$$

$$= \frac{\lambda_k}{p}(R\text{사용})$$

$$\underline{\mu} = \begin{bmatrix}10\\5\end{bmatrix} \Sigma = \begin{bmatrix}9 & 2\\2 & 4\end{bmatrix}$$

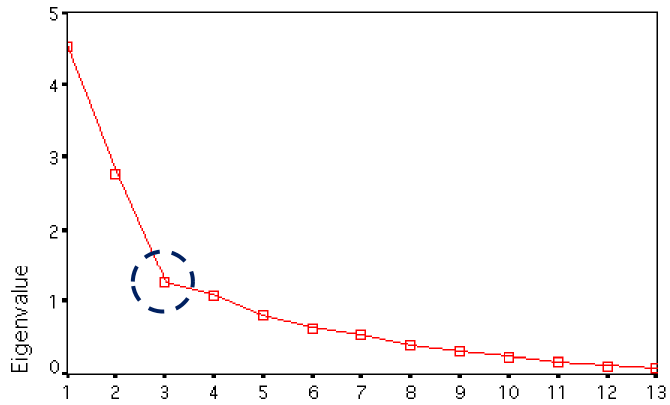


$$\lambda_1 = 9.7$$
$$\lambda_2 = 3.2$$

$$\begin{pmatrix}y_1\\y_2\end{pmatrix} = \begin{bmatrix}0.94 & 0.33\\-.33 & 0.94\end{bmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix}$$

- 주성분 개수 결정
  - 80% rule (공분산 행렬 사용 시)
  - 고유치 1 이상 (상관계수 행렬 사용 시)

  - Scree 도표 이용 (페이지 80)
    - Y축 고유치, x 축을 주성분 순차 번호 산점도
    - 고유치가 감소 경향을 시각적으로 표현
    - 급격히 감소하는 곳에서 주성분 개수 결정 (elbow)
    - 시각적 표현, 실제로는 80% 규칙 이용



선형계수 loading (l_{ij})

▪부하 (loading) 정의 (페이지 90)

- 공분산 행렬로부터 얻어지는 $\underline{l}_j = \sqrt{\lambda_j}\,\underline{e}_j$ 을 주성분 부하(component loading)라 정의
- 주성분 변수를 계산할 때 사용되는 선형계수 값

$$\begin{pmatrix} l_{11} & l_{12} & \cdots l_{1p} \\ l_{21} & l_{12} & \cdots l_{2p} \\ \vdots & & \\ l_{p1} & l_{p2} & \cdots l_{pp} \end{pmatrix} = \begin{pmatrix} \underline{e}_1^{'}(\text{고유치 } \lambda_1 \text{ 대응 고유벡터}) \\ \underline{e}_2^{'}(\text{고유치 } \lambda_2 \text{ 대응 고유벡터}) \\ \vdots \\ \underline{e}_p^{'}(\text{고유치 } \lambda_p \text{ 대응 고유벡터}) \end{pmatrix}$$

▪사용

- 주성분 변수를 계산할 때 원변수가 주성분 변수에 미치는 영향 정도가 부하이다.
- 부하 값이 크다는 것은 원변수의 영향력이 크다. (*전제조건: 원변수의 단위가 유사, 그렇지 않으면 상관계수 행렬 사용)
- 이를 이용하여 주성분 변수의 이름 부여한다.

▪주성분 성질2

- 주성분의 변동은 고유치와 같고, 변동의 크기는 제일, 제이, … 순이다. $Var(Y_i) = \underline{e}_i^{'}\Sigma\underline{e}_i = \lambda_i$
- 주성분 변수는 서로 독립이다.

$$Cov(Y_i, Y_k) = \underline{e}_i^{'}\Sigma\underline{e}_k = 0, \;\; for\ i \neq k$$

- 주성분 변수의 변동 합은 원변수 변동 합과 동일하다.

$$\sum_{i=1}^{p} V(Y_i) = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} V(X_i)$$

- 주성분 변수는 계산 식을 실제 데이터에 의해 계산된 값을 주성분 점수(score)라 한다. k번째 개체의 j번째 주성분 점수 계산 식은 다음과 같다.

$$y_{ki} = l_{i1}x_{k1} + l_{i2}x_{k2} + \ldots + l_{ip}x_{kp}$$

▪주성분 계수 시각적 표현 (페이지 95)

- 주성분 변수(점수) 이름 부여에 도움
- 실제 계수 값을 보고도 판단할 수 있으나 시각적 표현으로 인하여 손쉽게 이름 부여 가능

주성분 설명 비율, 계수 및 이름 부여 APPLICANT.TXT

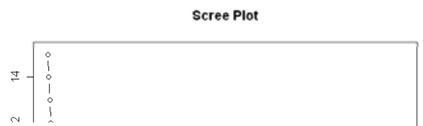
▪ APPLICANT.TXT

```
pca_app=prcomp(app_s        )
summary(pca_app)
```

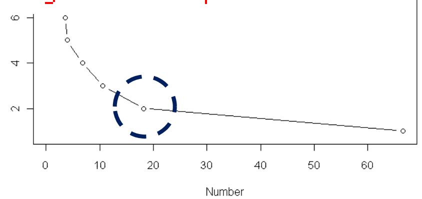
```
> summary(pca_app)
Importance of components:
                          PC1   PC2    PC3    PC4
Standard deviation      8.157 4.264 3.2544 2.6015
Proportion of Variance  0.543 0.148 0.0864 0.0552
Cumulative Proportion   0.543 0.691 0.7778 0.8330
```

- STDEV의 제곱이 고유치에 해당함.
- 페이지 86의 결과와 비교해 보자
- 리커트 척도와 같은 순서형 변수들은 차원 축약이 잘 되지 않음 (why? 변수 상관 계수 낮음)

▪ Scree plot 그리기




```
> eigenv=pca_app$sdev^2
> x=seq(1:15)
> plot(x,eigenv,main="Scree plot",type="b",
+ xlab="Number", ylab="Eigen Value")
```



- 이상하다
- 4가 아니네… T.T
- 그냥 시각적 표현?

```
> names(pca_app)
[1] "sdev"      "rotation" "center"    "scale"     "x"
```

▪ 계수 출력

```
> round(pca_app$rotation[,1:4],3)
            PC1    PC2    PC3    PC4
X1.FL.   -0.149 -0.371  0.200  0.277
X2.APP.  -0.132  0.029  0.042 -0.134
X3.AA.   -0.030 -0.102 -0.131 -0.603
X4.LA.   -0.203  0.093  0.620 -0.126
X5.SC.   -0.231  0.236 -0.189  0.072
X6.LC.   -0.337  0.196 -0.125 -0.053
X7.HON.  -0.120  0.301  0.447 -0.256
X8.SMS.  -0.379  0.090 -0.282  0.172
X9.EXP.  -0.164 -0.636  0.025 -0.166
X10.DRV. -0.316  0.012 -0.113  0.135
X11.AMB. -0.312  0.122 -0.245  0.147
X12.GSP. -0.339  0.074 -0.050 -0.206
X13.POT. -0.357  0.025  0.041 -0.317
X14.KJ.  -0.226  0.045  0.385  0.460
X15.SUIT. -0.274 -0.471  0.017  0.016
```

- 고유치는 무수히 많이 존재, 부호가 반대(? 페이지 87).
- 각 열의 제곱 합은 1이다. $e_i'e_i=1$
- 각 열끼리 내적 합은 0이다. $e_i'e_j=0$

```
> plot(pca_app,main="Scree plot",type="l")
```

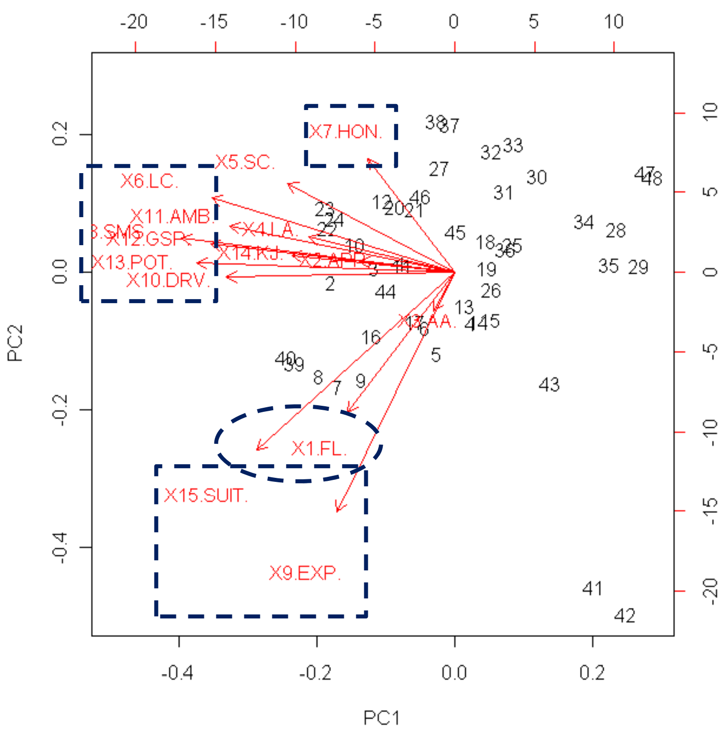
- 위 plot() 함수로도 Scree plot을 그릴 수 있음

http://wolfpack.hnu.ac.kr

주성분 설명 비율, 계수 및 이름 부여 APPLICANT.TXT (계속)

Biplot 결과

```
> biplot(pca_app)
```




```
biplot(pca_app,choices=1:2)
```

- 2:3으로 하면 2와 3이 그려짐

이름 부여 (페이지 93)

- 다소 주관적, 상대적 크기

- 제일 주성분: 정신적 지적 능력
- 제이 주성분:경험
- 제삼 주성분: 심성
- 제사 주성분:학교성적

주성분 점수

```
> round(pca_app$x[,1:4],2)
        PC1   PC2   PC3   PC4
[1,]  -4.30  0.38 -1.76  5.62
[2,] -10.14 -0.42  0.09  3.09
[3,]  -6.53  0.17 -0.32  4.53
[4,]   1.33 -2.18  1.10 -2.81
[5,]  -1.48 -3.49  3.25 -2.45
[6,]  -2.38 -2.42  1.04 -1.18
[7,]  -9.59 -4.92  0.31  0.10
[8,] -11.07 -4.49  0.08 -0.10
[9,]  -7.59 -4.60  1.46  0.52
```

http://wolfpack.hnu.ac.kr

주성분 설명 비율, 계수 및 이름부여 BIG8.TXT

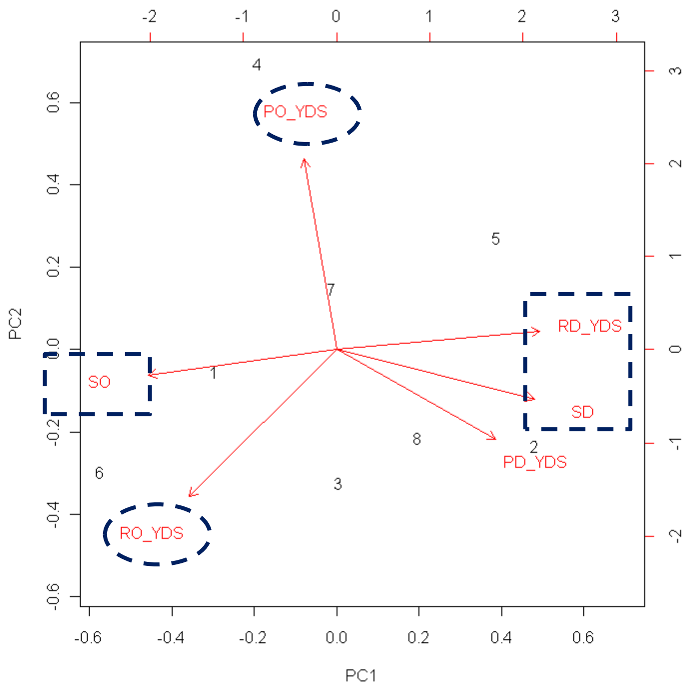
- **BIG8.TXT**
  - 상관계수 행렬을 사용할 경우에는 scale 옵션 사용

```
pca_big=prcomp(Big8s,scale=T)
summary(pca_big)
eigenv=pca_big$sdev^2
x=seq(1:6)
plot(x,eigenv,main="Scree plot",type="b",
xlab="Number", ylab="Eigen Value")
round(pca_big$rotation[,1:4],3)
biplot(pca_big)
```

  - 측정형 변수들이라 2개 주성분만으로 80% 이상

```
> round(pca_big$rotation[,1:4],3)
          PC1    PC2   PC3    PC4
RO_YDS -0.365 -0.556 0.156 -0.082
RD_YDS  0.502  0.070 0.012 -0.622
PO_YDS -0.080  0.725 0.529  0.079
PD_YDS  0.393 -0.341 0.628  0.485
SO     -0.466 -0.095 0.519 -0.525
SD      0.487 -0.187 0.180 -0.300
```

```
> summary(pca_big)
Importance of components:
                         PC1   PC2    PC3    PC4
Standard deviation     1.923 1.248 0.7332 0.3500
Proportion of Variance 0.616 0.260 0.0896 0.0204
Cumulative Proportion  0.616 0.876 0.9658 0.9862
```



- 주성분 이름
  - 제1주성분: 작을수록 수비능력 높음 (-)
  - 제2: Passing 공격 능력(+)

주성분 점수 활용

- 주성분 점수 principal component score (페이지 97)
  - 고유 벡터가 loading 벡터(선형계수 벡터)이다.
  - x_i, y_i 들은 변수를 나타낸다.

$$\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_p \end{pmatrix} = \begin{pmatrix} l_{11}\ l_{12} \cdots l_{1p} \\ l_{21}\ l_{22} \cdots l_{2p} \\ \cdots \\ l_{p1}\ l_{p2} \cdots l_{pp} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_p \end{pmatrix} = \begin{pmatrix} e_{11}\ e_{12} \cdots e_{1p} \\ e_{21}\ e_{22} \cdots e_{2p} \\ \cdots \\ e_{p1}\ e_{p2} \cdots e_{pp} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_p \end{pmatrix}$$

- 활용
  - Anomaly 진단
    - 주요 성분 활용: 95% CI 벗어나는 개체
      - k=적절한 주성분 개수

$$\frac{y_1}{\lambda_1} + \frac{y_2}{\lambda_2} + \ldots \frac{y_k}{\lambda_k} \le \chi^2(df = k)$$

    - 잔차 주성분에서 CI 벗어나는 개체

$$\frac{y_{k+1}}{\lambda_{k+1}} + \frac{y_{k+2}}{\lambda_{k+2}} + \ldots \frac{y_p}{\lambda_p} \le \chi^2(df = p-k+1)$$

- 데이터 스크린
  - 주성분 변수에 의해 축약된 단일 개념 측정
- 이상치 발견
  - 각 주성분 일변량 분석: 상자 수염 그림
  - 주성분들의 산점도 활용 (변수들간의 관계 속에서 이상치)
- 회귀분석: 설명변수의 다중 공선성 문제 해결
  - 설명변수의 상관행렬(공분산행렬)로부터 구한 주성분 점수를 설명변수로 이용
- 원 데이터 다변량 정규분석 검정
  - 주성분의 일변량 정규성 검정

주성분 점수 활용 (APPLICANT.TXT)

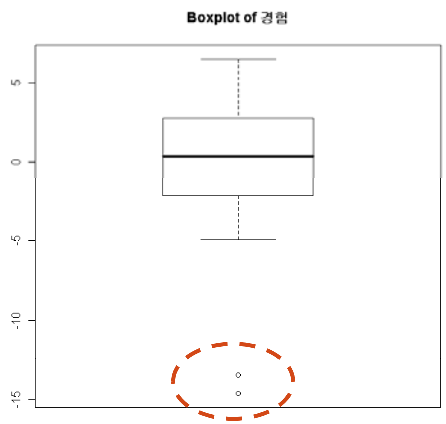
- APPLICANT.TXT
  - 회귀분석: 설명변수의 다중 공선성 문제 해결
    - 15개 능력 측정 설명변수 => 15개 주성분 설명변수화:
      - 여기서는 15개 주성분 모두 사용하여 유의성 검정, 일반적으로 유의한 주성분 설명변수 개수와 80% 규칙에 의한 주성분 개수와 유사
  - 데이터 스크린
    - 주성분 변수(단일 지표)에 의한 개체 순위 및 이상치 발견
    - 주성분 변수들의 산점도 활용: 이상치 진단

```
pca_s=data.frame(round(pca_app$x[,1:4],3))
names(pca_s)
attach(pca_s)
boxplot(PC1,main="Boxplot of 지적")
boxplot(PC2,main="Boxplot of 경험")
boxplot(PC3,main="Boxplot of 심성")
plot(PC2,PC1,main="Scatter plot of (지적, 경험)")
```

- 경험 점수에서 이상치 존재, 표시되면 좋은데
- 산점도: 위와 같이 하면 개체 ID가 표시되지 않는다.

- loading 값 및 주성분 이름

	지적 (-)	경험(-)	심성(+)	학교성적(-)
	PC1	PC2	PC3	PC4
X1.FL.	-0.149	-0.371	0.200	0.277
X2.APP.	-0.132	0.029	0.042	-0.134
X3.AA.	-0.030	-0.102	-0.131	-0.603
X4.LA.	-0.203	0.093	0.620	-0.126
X5.SC.	-0.231	0.236	-0.189	0.072
X6.LC.	-0.337	0.196	-0.125	-0.053
X7.HON.	-0.120	0.301	0.447	-0.256
X8.SMS.	-0.379	0.090	-0.282	0.172
X9.EXP.	-0.164	-0.636	0.025	-0.166
X10.DRV.	-0.316	-0.012	-0.113	0.135
X11.AMB.	-0.312	0.122	-0.245	0.147
X12.GSP.	-0.339	0.074	-0.050	-0.206
X13.POT.	-0.357	0.025	0.041	-0.317
X14.KJ.	-0.226	0.045	0.385	0.460
X15.SUIT.	-0.274	-0.471	0.017	0.016

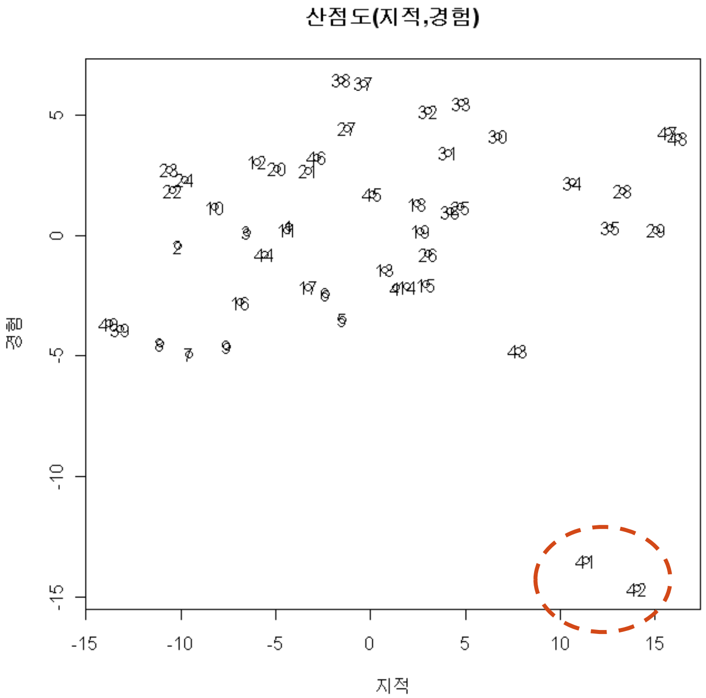


주성분 점수 활용 (APPLICANT.TXT)

개체 ID가 표현된 산점도

```
id=app[,1:1]
all=data.frame(cbind(id,pca_app$x[,1:4]))
names(all)
attach(all)
plot(PC1,PC2,main="산점도(지적,경험)",
xlab="지적",ylab="경험")
text(PC1,PC2,labels=as.character(id))
```

- 개체 (40, 41) 경험 능력 만땅, 지적 능력 매우 낮음



주성분 점수에 의한 순위

- 경험: 점수 낮을수록 경험 높은 지원자

```
> #경험 점수에 의한 SORT
> all_sort=all[order(PC2),]
> all_sort
   id        PC1         PC2         PC3
42 42  14.0307125 -14.6597223 -3.52656475
41 41  11.3814069 -13.5076309 -3.03192993
7   7  -9.5935187  -4.9222794  0.30695664
43 43   7.8022649  -4.7751857  1.28965783
9   9  -7.5935693  -4.6013232  1.46089489
8   8 -11.0723255  -4.4857087  0.08333453
```

- order(PC2,PC1) ➔ 2단계 정렬

- 학점: 점수 높을수록 학점이 높은 지원자

```
> #대학 학점에 의한 SORT
> all_sort=all[order(PC4,decreasing=T),]
> round(all_sort,2)
   id   PC1   PC2   PC3  PC4
29 29 15.05  0.26 -2.52 6.49
1   1 -4.30  0.38 -1.76 5.62
28 28 13.28  1.87 -0.41 4.87
3   3 -6.53  0.17 -0.32 4.53
```

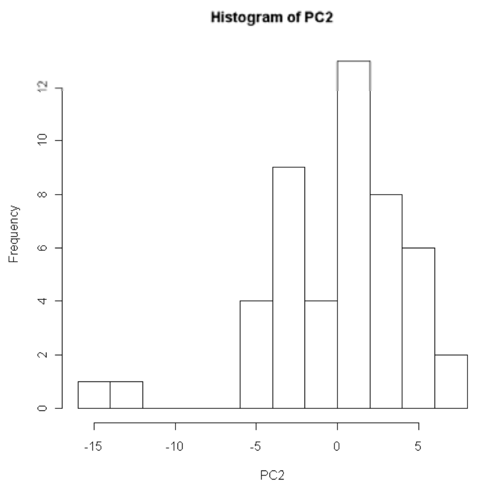
주성분 점수 활용 (APPLICANT.TXT)

- 원 데이터 다변량 정규분석 검정
  - 주성분의 일변량 정규성 검정 (경험 주성분)

```
> hist(all$PC2,nclass=10)
> shapiro.test(all$PC2)

        Shapiro-Wilk normality test

data:  all$PC2
W = 0.8972, p-value = 0.0005118
```



  - 정규성이 무너진 이유는 이상치가 존재하기 때문으로 보인다.

```
> all.s=subset(all,PC2>-10)
> shapiro.test(all.s$PC2)

        Shapiro-Wilk normality test

data:  all.s$PC2
W = 0.9723, p-value = 0.3357
```

- Anomaly 진단

```
attach(pca_app)
nm=PC1^2/sdev[1:1]^2+PC2^2/sdev[2:2]^2+
PC3^2/sdev[3:3]^2+PC4^2/sdev[4:4]^2
cat("Rejection Value",qchisq(0.95,4),"\n")
for(i in 1:length(nm)){
flag[i]=c("A")
if(nm[i]<qchisq(0.95,4)){flag[i]=c("O")}
cat(i,nm[i],flag[i],"\n")
}
```

```
> cat("Rejection Value",qchisq(0.95,4),"\n"
Rejection Value 9.487729
> for(i in 1:length(nm)){
+ flag[i]=c("A")
+ if(nm[i]<qchisq(0.95,4)){flag[i]=c("O")}
+ cat(i,nm[i],flag[i],"\n")
+ }
1 10.72982 A
2 8.45574 O
3 9.1769 O
4 6.811263 O
```

첫 번째 지원자는 다른 개체와 다른 능력 anomaly

주성분 점수 활용 (Big8.TXT)

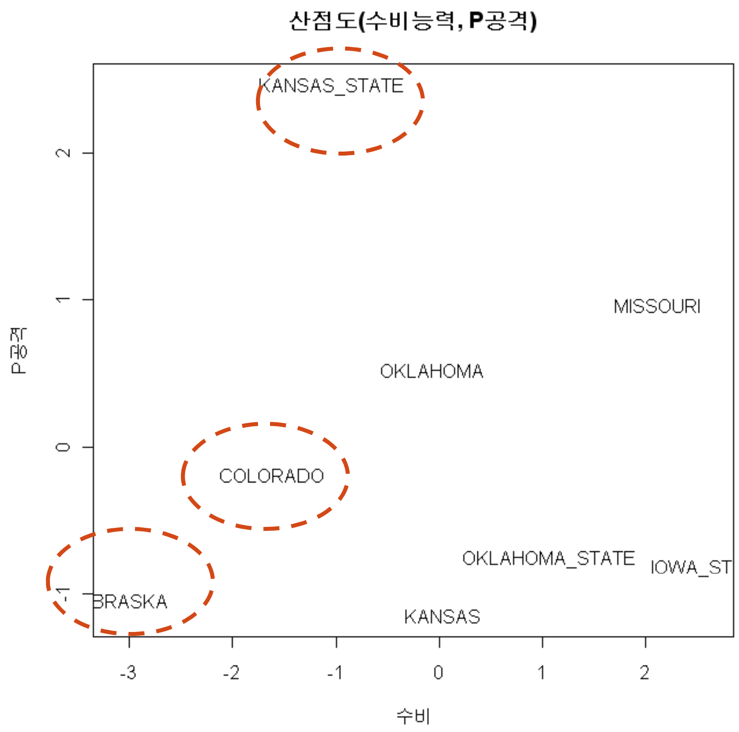
- BIG8.TXT
  - 데이터 스크린
    - 주성분 변수(단일 지표)에 의한 개체 순위 및 이상치 발견
    - 주성분 변수들의 산점도 활용: 이상치 진단

```
> big8_a=list(Name=Big8[,1:1],Winpct=Big8[,13:13],
+ pca1=pca_big$x[,1:1],pca2=pca_big$x[,2:2])
> big8_zz=data.frame(big8_a)
> big8_zz
           Name    Winpct        pca1       pca2
1      COLORADO 0.9090909 -1.61446496 -0.1849591
2    IOWA_STATE 0.0000000  2.61875501 -0.8195749
3        KANSAS 0.5454545  0.03372645 -1.1427498
4  KANSAS_STATE 0.8181818 -1.04059015  2.4598006
5      MISSOURI 0.2916667  2.11863748  0.9611206
6      NEBRASKA 1.0000000 -3.11658333 -1.0418036
7      OKLAHOMA 0.5454545 -0.06599738  0.5209985
8 OKLAHOMA STATE 0.3181818  1.06651689 -0.7528323
```

```
plot(big8_zz$pca1,big8_zz$pca2,
main="산점도(수비능력, P공격)",
xlab="수비",ylab="P공격",type="n")
text(big8_zz$pca1,big8_zz$pca2,labels=Name)
```

- KANSAS State: Passing 공격 강함, 승률 3위
- Nebraska: 수비능력 강함, 승률 1위
- Colorado: 수비 2위, P 공격 4위, 승률 2위

- loading 값 및 주성분 이름 슬라이드 36
  - 제1주성분: 수비능력(-)
  - 제2주성분: 패싱 공격능력(+)

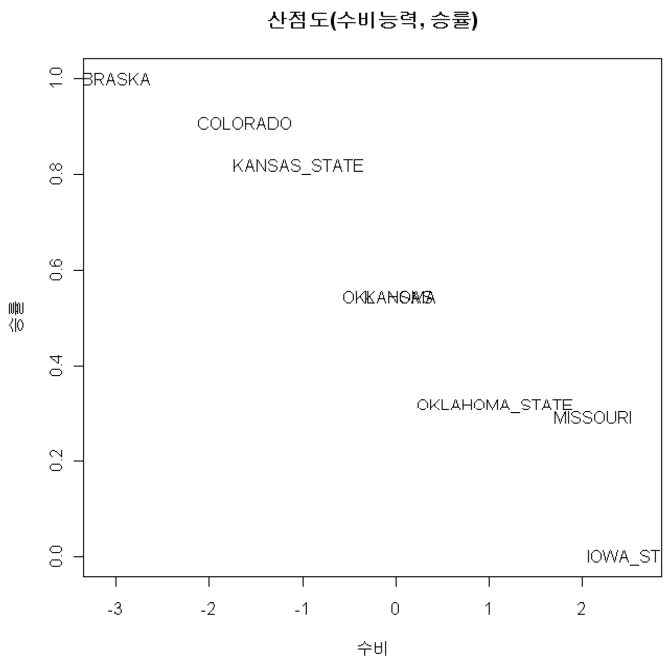


http://wolfpack.hnu.ac.kr

주성분 점수 활용 2 (Big8.TXT)

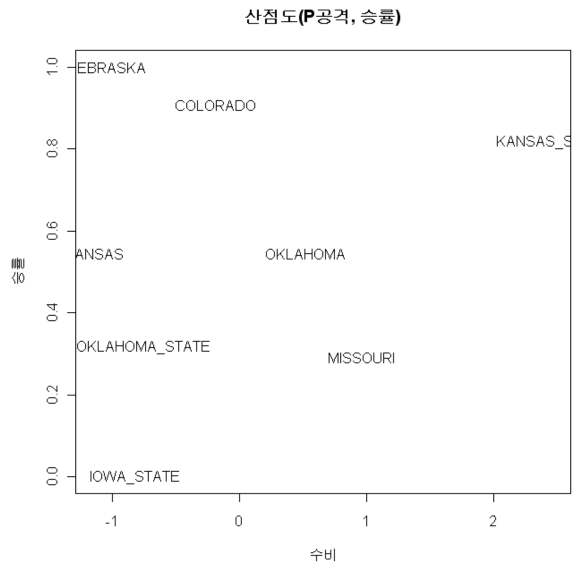
- 승률과 수비능력 주성분 비교
  - 상관관계 매우 높음, 수비 능력이 높을수록(낮은 값이 능력 좋음) 승률 높아진다.

```
> cor(big8_zz$pca1,big8_zz$Winpct)
[1] -0.972027
> plot(big8_zz$pca1,big8_zz$pca2,
+ main="산점도(수비능력, P공격)",
+ xlab="수비",ylab="P공격",type="n")
> text(big8_zz$pca1,big8_zz$pca2,labels=Name)
```



- 승률과 Passing 공격능력 비교
  - P. 능력은 승률과 다소 무관
  - 수비능력에 의해 승률이 좌우됨.

```
> cor(big8_zz$pca2,big8_zz$Winpct)
[1] 0.1679537
> plot(big8_zz$pca2,big8_zz$Winpct,
+ main="산점도(P공격, 승률)",
+ xlab="수비",ylab="승률",type="n")
> text(big8_zz$pca2,big8_zz$Winpct,labels=Name)
```

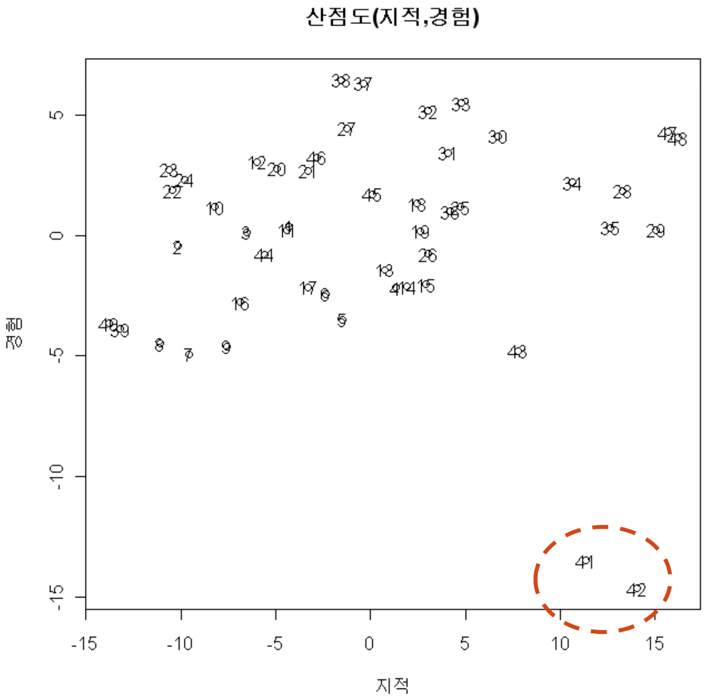


주성분 점수 활용 (APPLICANT.TXT)

개체 ID가 표현된 산점도

```
id=app[,1:1]
all=data.frame(cbind(id,pca_app$x[,1:4]))
names(all)
attach(all)
plot(PC1,PC2,main="산점도(지적,경험)",
xlab="지적",ylab="경험")
text(PC1,PC2,labels=as.character(id))
```

- 개체 (40, 41) 경험 능력 만땅, 지적 능력 매우 낮음



주성분 점수에 의한 순위

- 경험: 점수 낮을수록 경험 높은 지원자

```
> #경험 점수에 의한 SORT
> all_sort=all[order(PC2),]
> all_sort
   id        PC1         PC2         PC3
42 42  14.0307125 -14.6597223 -3.52656475
41 41  11.3814069 -13.5076309 -3.03192993
7   7  -9.5935187  -4.9222794  0.30695664
43 43   7.8022649  -4.7751857  1.28965783
9   9  -7.5935693  -4.6013232  1.46089489
8   8 -11.0723255  -4.4857087  0.08333453
```

- order(PC2,PC1) ➔ 2단계 정렬

- 학점: 점수 높을수록 학점이 높은 지원자

```
> #대학 학점에 의한 SORT
> all_sort=all[order(PC4,decreasing=T),]
> round(all_sort,2)
   id   PC1   PC2   PC3  PC4
29 29 15.05  0.26 -2.52 6.49
1   1 -4.30  0.38 -1.76 5.62
28 28 13.28  1.87 -0.41 4.87
3   3 -6.53  0.17 -0.32 4.53
```

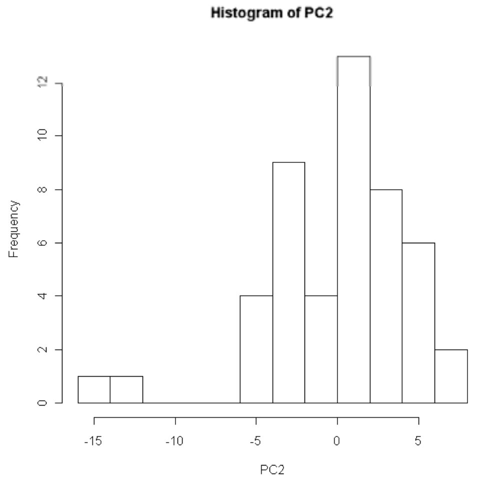
주성분 점수 활용 (APPLICANT.TXT)

- 원 데이터 다변량 정규분석 검정
  - 주성분의 일변량 정규성 검정 (경험 주성분)

```
> hist(all$PC2,nclass=10)
> shapiro.test(all$PC2)

        Shapiro-Wilk normality test

data:  all$PC2
W = 0.8972, p-value = 0.0005118
```



  - 정규성이 무너진 이유는 이상치가 존재하기 때문으로 보인다.

```
> all.s=subset(all,PC2>-10)
> shapiro.test(all.s$PC2)

        Shapiro-Wilk normality test

data:  all.s$PC2
W = 0.9723, p-value = 0.3357
```

- Anomaly 진단

```
attach(pca_app)
nm=PC1^2/sdev[1:1]^2+PC2^2/sdev[2:2]^2+
PC3^2/sdev[3:3]^2+PC4^2/sdev[4:4]^2
cat("Rejection Value",qchisq(0.95,4),"\n")
for(i in 1:length(nm)){
flag[i]=c("A")
if(nm[i]<qchisq(0.95,4)){flag[i]=c("O")}
cat(i,nm[i],flag[i],"\n")
}
```

```
> cat("Rejection Value",qchisq(0.95,4),"\n"
Rejection Value 9.487729
> for(i in 1:length(nm)){
+ flag[i]=c("A")
+ if(nm[i]<qchisq(0.95,4)){flag[i]=c("O")}
+ cat(i,nm[i],flag[i],"\n")
+ }
1 10.72982 A
2 8.45574 O
3 9.1769 O
4 6.811263 O
```

첫 번째 지원자는 다른 개체와 다른 능력 anomaly