**HW 3 for Multivariate Statistics (I)**

**Chapter 3. Factor Analysis**

**학과 : 전기공학과**

**학번 : 201724570**

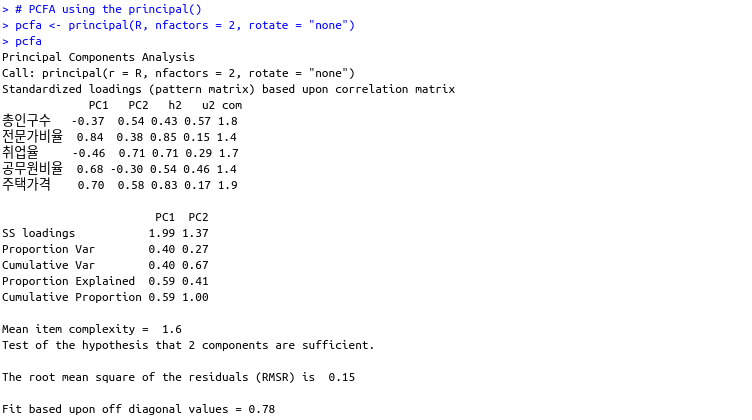
**이름 : 정석규**

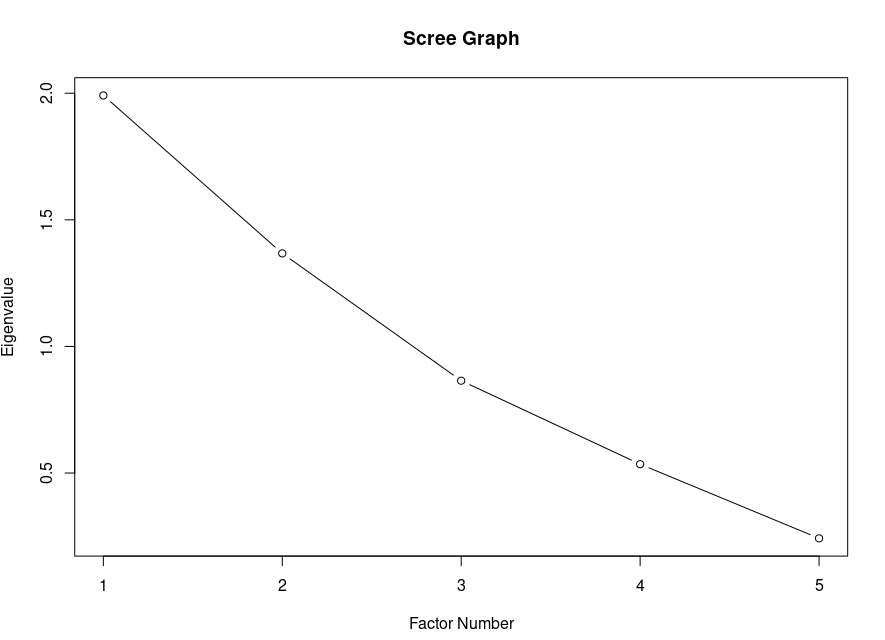
Consider the Exercise 3.5 in page 190.

Data(censustract.txt) in Exercise 2.7 is collected by 5 measurement variables(total population(per thousand), ratio of experts(%), employment rate for people over 16 years of age(%), the ratio of civil servants(%), and house prices(per 100 milliion won)) in 61 regions.

For this data, solve the problems (1) ~ (5) described in Exercise 3.7.

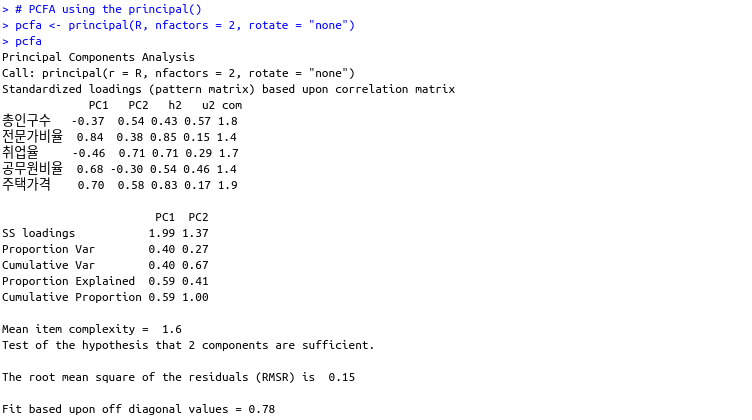
**(1) PCFA를 실시하여 스크리그림을 통하여 인자개수를 정하고 총기여율을 구하라**

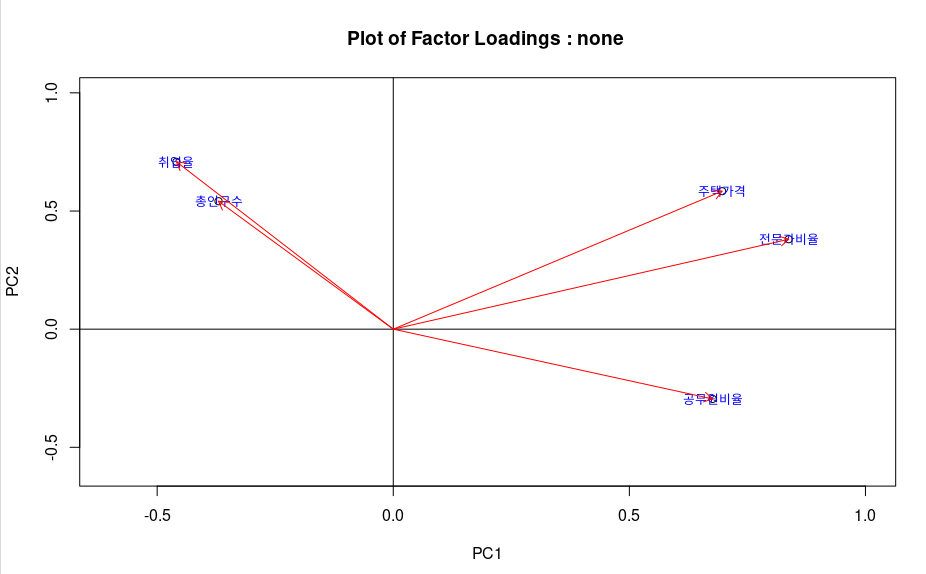




고유근이 1이상인 경우까지만 인자수로 채택한다. 해당 그래프에서는 2까지만 1이 넘기 때문에 인자개수를 2개로 정한다. Proportion Var은 PC1, PC2에 대하여 각각 40%, 27%로, 총 기여율은 67%라고 할 수 있다.

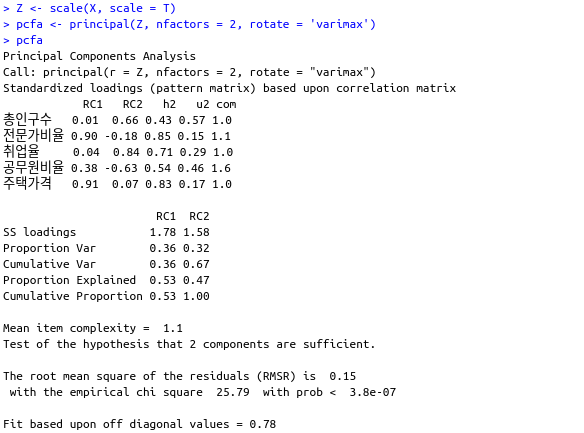
**(2) 인자적재값과 인자적재그림을 통하여 인자를 해석하라**

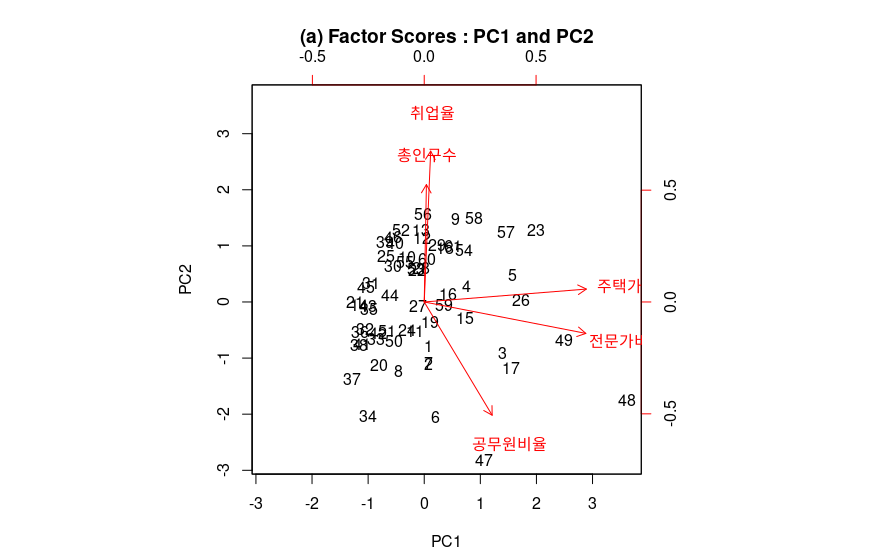




첫번째 요인에 대해 총 인구수와 취업은 음의 값을 가지고, 전문가 비율과 공무원 비율, 주택 가격은 양의 값을 가지고 있음을 확인할 수 있다. 더불어 첫번째 요인의 값은 전체적으로 절댓값의 크기가 크고, 전문직 비율과 주택 가격이 높은 결과 수도권과 도시 인근의 행정구역에 대한 정보를 저장하고 있음을 추측할 수 있다. 인자 적재 그림을 확인한 결과 취업율, 총인구수는 공무원 비율과 다른 성격을 가지는 변수라고 생각할 수 있다. 또한 PC1 축에 대해서 전문가의 비율과 공무원의 비율의 각이 작기 때문에 두 변수가 PC1을 결정하는 변수라고 생각할 수 있다.

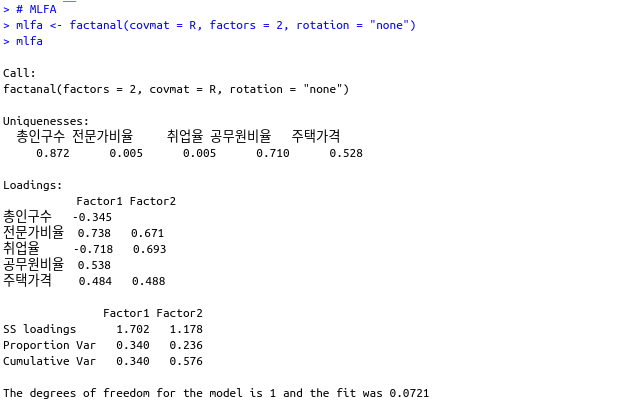
**(3) 인자점수그림을 통해 형성된 군집의 특성을 살펴보라**

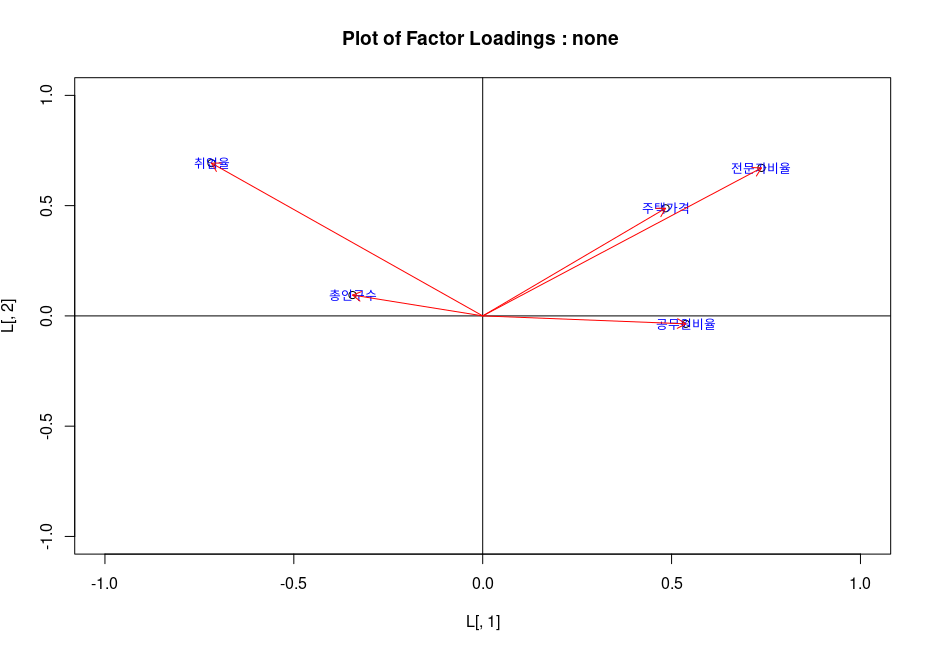


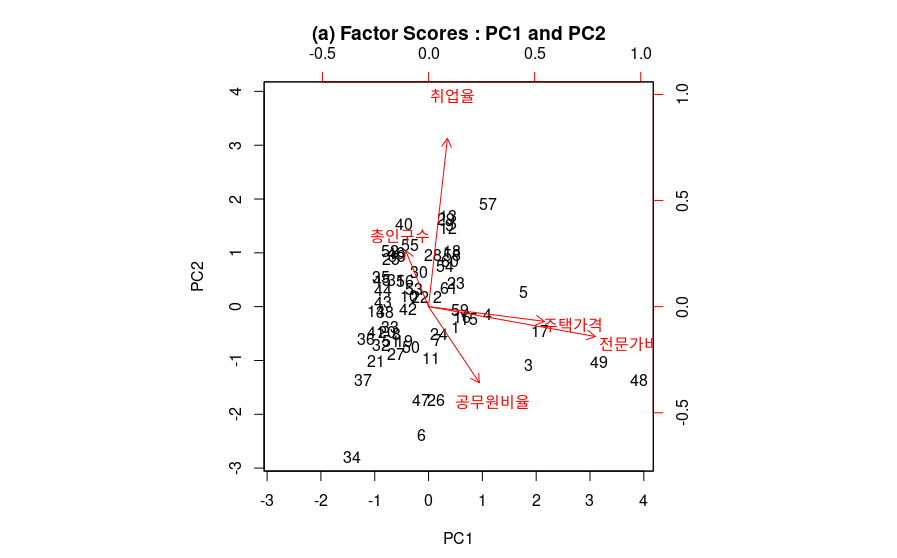


직교회전(varimax rotation)을 하고 PCFA를 실행하고 binplot을 통해 인자 적재 그림과 인자 점수 그림을 표현한 결과, 52, 56, 9, 58번과 같은 행정구역은 취업율과 총 인구수가 높은 행정구역임을 알 수 있다. 마찬가지로 4사분면에 위치한 6, 17, 47, 48, 49 행정구역은 전문가와 공무원의 비율이 높은 집중된 도시 구역임을 알 수 있다.

**(4) (1)의 인자개수에 대해 MLFA를 실시하고 (2)~(3)을 시행한 후에 결과를 서로 비교하라**



 MLFA를 시행한 결과, 첫번째 요인에 대해서 PCFA와 비슷한 인자 적재값을 확인하고 있음을 확인할 수 있다. 총 기여율은 요인1과 요인2가 각각 34%, 23%를 더한 57%로 PCFA에 비해 낮은 총 기여율을 가지고 있음을 알 수 있다. 인자 적재 그림을 확인한 결과 PCFA와는 다르게 요인 1에 대해서 총 인구수와 공무원의 비율이 높게 결정하고 있음을 알 수 있다.



직교회전(varimax rotation)을 하고 MLFA를 수행한 결과를 binplot에 적용한 결과 생성되는 그룹에 대해서 큰 차이가 없음을 확인할 수 있었다.

**Code works**

# Problem (1) ----------------------------------------------------------------------------------------------------------

# Data Matrix X

df\_3\_5 = read.table("Multivariate\_Statistics/Datasets/censustract.txt", fileEncoding = "EUC-KR", header = T)

X = df\_3\_5

rownames <- rownames(X)

colnames <- colnames(X)

# num cols

p = ncol(X)

n = nrow(X)

# Correlation Matrix

R = round(cor(X), 3)

R

# Spectral Decomposition

eigen.R = eigen(R)

round(eigen.R$values, 2)

V = round(eigen.R$vectors, 2)

# Number of factors : m

gof = eigen.R$values/p\*100

round(gof,3)

plot(eigen.R$values, type = 'b', main = 'Scree Graph', xlab = "Factor Number", ylab = "Eigenvalue")

# Problem(2) ----------------------------------------------------------------------------------------------------------------

# Factor Loadings and Communality

V2 = V[,1:2]

L = V2%\*%diag(sqrt(eigen.R$values[1:2]))

rownames(L) <- colnames(X)

colnames(L) <- c("PC1", "PC2")

round(L, 3)

round(diag(L%\*%t(L)), 3)

# Specific Variance : Psi

Psi = diag(R-L%\*%t(L))

round(Psi, 3)

# Residual Matrix

Rm = R-(L%\*%t(L) + diag(Psi))

round(Rm, 3)

# PCFA using the principal()

library(psych)

pcfa <- principal(R, nfactors = 2, rotate = "none")

pcfa

round(pcfa$values, 2)

gof = pcfa$values/p\*100

round(gof, 3)

round(pcfa$residual, 2)

# Factor Loadings Plot

L = pcfa$loadings[,1:2]

rownames(L) <- colnames(X)

colnames(L) <- c("PC1", "PC2")

lim <- range(pretty(L))

plot(L[,1], L[,2], main = "Plot of Factor Loadings : none ", xlab = "PC1", ylab = "PC2", xlim=lim, ylim=lim)

text(L[,1], L[,2], labels = rownames(L), cex = 0.8, col = "blue", abline(v=0, h=0))

arrows(0,0,L[,1],L[,2], col=2, code=2, length = 0.1)

round(pcfa$values, 3)

gof=pcfa$values/p\*100 # Goodness-of fit

round(gof, 3)

# Problem(3) ---------------------------------------------------------------------------------------------------------

# Factor Score

library(psych)

Z <- scale(X, scale = T)

pcfa <- principal(Z, nfactors = 2, rotate = 'varimax')

pcfa

# Residual Matrix

L=pcfa$loading[, 1:2]

round(L, 3)

Psi=pcfa$uniquenesses

round(Psi, 3)

Rm = R-(L%\*%t(L) + diag(Psi))

round(Rm, 3)

lim<-range(pretty(L))

plot(L[,1], L[,2],main="(a) PC Factor Loadings : PC1 and PC2", xlab="PC1", ylab="PC2",

xlim=lim, ylim=lim)

text(L[,1], L[, 2], labels=rownames(L), cex=0.8, col="blue", pos=1)

abline(v=0, h=0)

arrows(0,0, L[,1], L[, 2], col=2, code=2, length=0.1)

fpc=pcfa$scores

round(fpc, 3)

lim<-range(pretty(fpc))

plot(fpc[,1], fpc[,2],main="(a) Factor Scores : PC1 and PC2", xlab="PC1", ylab="PC2",

xlim=lim, ylim=lim)

text(fpc[,1], fpc[,2], labels=rownames(fpc), cex=0.8, col="blue", pos=1)

abline(v=0, h=0)

biplot(pcfa$scores[,1:2], pcfa$loadings[,1:2], main="(a) Factor Scores : PC1 and PC2", xlab="PC1", ylab="PC2")

# Problem(4) ----------------------------------------------------------------------------------------------------

# MLFA

mlfa <- factanal(covmat = R, factors = 2, rotation = "none")

mlfa

# Residual Matrix

L = mlfa$loadings[,1:2]

Psi = mlfa$uniquenesses

Rm = R-(L%\*%t(L) + diag(Psi))

round(Rm, 3)

# Factor Loadings Plot

lim <- range(pretty(L))

plot(L[,1], L[,2], main = "Plot of Factor Loadings : none ", xlim=lim, ylim=lim)

text(L[,1], L[,2], labels = rownames(L), cex = 0.8, col = "blue", abline(v=0, h=0))

arrows(0,0,L[,1],L[,2], col=2, code=2, length = 0.1)

library(psych)

Z <- scale(X, scale = T)

mlfa <- factanal(Z, factors = 2, rotation = 'varimax', score = "regression")

mlfa

# Residual Matrix

L=mlfa$loading[, 1:2]

round(L, 3)

Psi=mlfa$uniquenesses

round(Psi, 3)

Rm = R-(L%\*%t(L) + diag(Psi))

round(Rm, 3)

lim<-range(pretty(L))

plot(L[,1], L[,2],main="(a) ML Factor Loadings : PC1 and PC2", xlab="PC1", ylab="PC2",

xlim=lim, ylim=lim)

text(L[,1], L[, 2], labels=rownames(L), cex=0.8, col="blue", pos=1)

abline(v=0, h=0)

arrows(0,0, L[,1], L[, 2], col=2, code=2, length=0.1)

fml=mlfa$scores

round(fml, 3)

lim<-range(pretty(fml))

plot(fml[,1], fml[,2],main="(a) Factor Scores : PC1 and PC2", xlab="PC1", ylab="PC2",

xlim=lim, ylim=lim)

text(fml[,1], fml[,2], labels=rownames(fml), cex=0.8, col="blue", pos=1)

abline(v=0, h=0)

biplot(mlfa$scores[,1:2], mlfa$loadings[,1:2], main="(a) Factor Scores : PC1 and PC2", xlab="PC1", ylab="PC2")