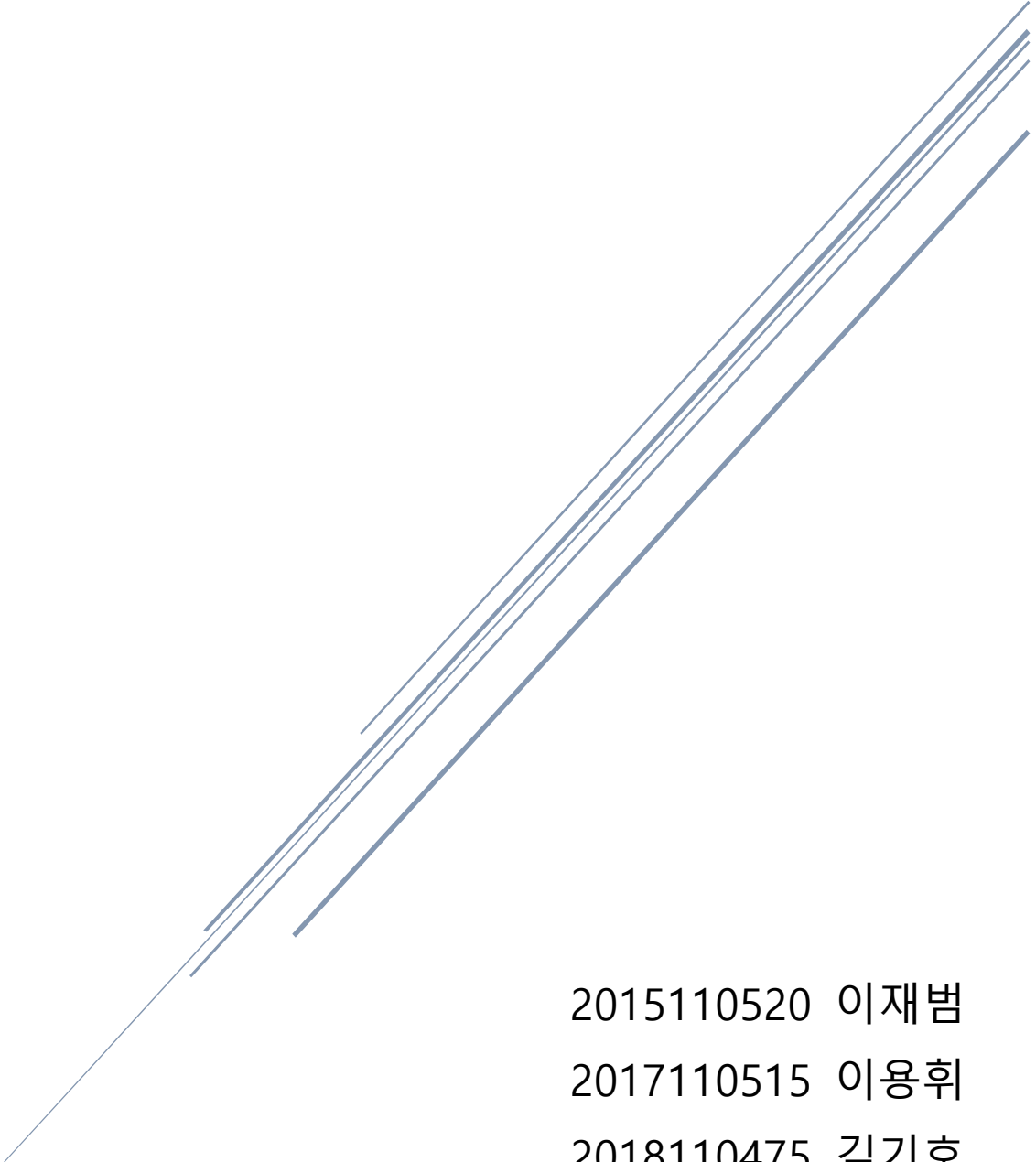


# 체지방 및 신체 둘레를 활용한 연령대별 신체 특성 분석



2015110520 이재범  
2017110515 이용휘  
2018110475 김기호  
2018110503 구현우

# 목차

## I. 서론

- 자료 배경 및 설명
- 분석 필요성 및 목적

## II. 본론

- 데이터 전처리
- 주성분 분석(Principal Component Analysis)
- MANOVA(Multivariate Analyses of Variance)

## III. 결론

## IV. 부록

# I. 서론

## - 자료 배경 및 설명

분석에 활용할 자료는 정확한 체지방을 측정하는 것은 불편하고 비용이 많이 들기 때문에 신체 둘레를 이용하여 체지방을 쉽게 예측해 보고자 수집된 자료이다. 자료는 252개의 row, 15개의 column으로 이루어져 있다.

<자료 설명>

Column	Details
Density	체중(gm)/체표면적( $\text{cm}^2$ ) 수중계량한 신체밀도
Bodyfat	체지방률(%)
Age	나이(year)
Weight	몸무게(pound)
Height	키(inches)
Neck	목의 둘레(cm)
Chest	가슴 둘레(cm)
Abdomen	복부 둘레(cm)
Hip	엉덩이 둘레(cm)
Thigh	허벅지 둘레(cm)
Knee	무릎 둘레(cm)
Ankle	발목 둘레(cm)
Biceps	이두박근 둘레(cm)
Forearm	전완근 둘레(cm)
Wrist	손목 둘레(cm)

## - 분석 필요성 및 목적

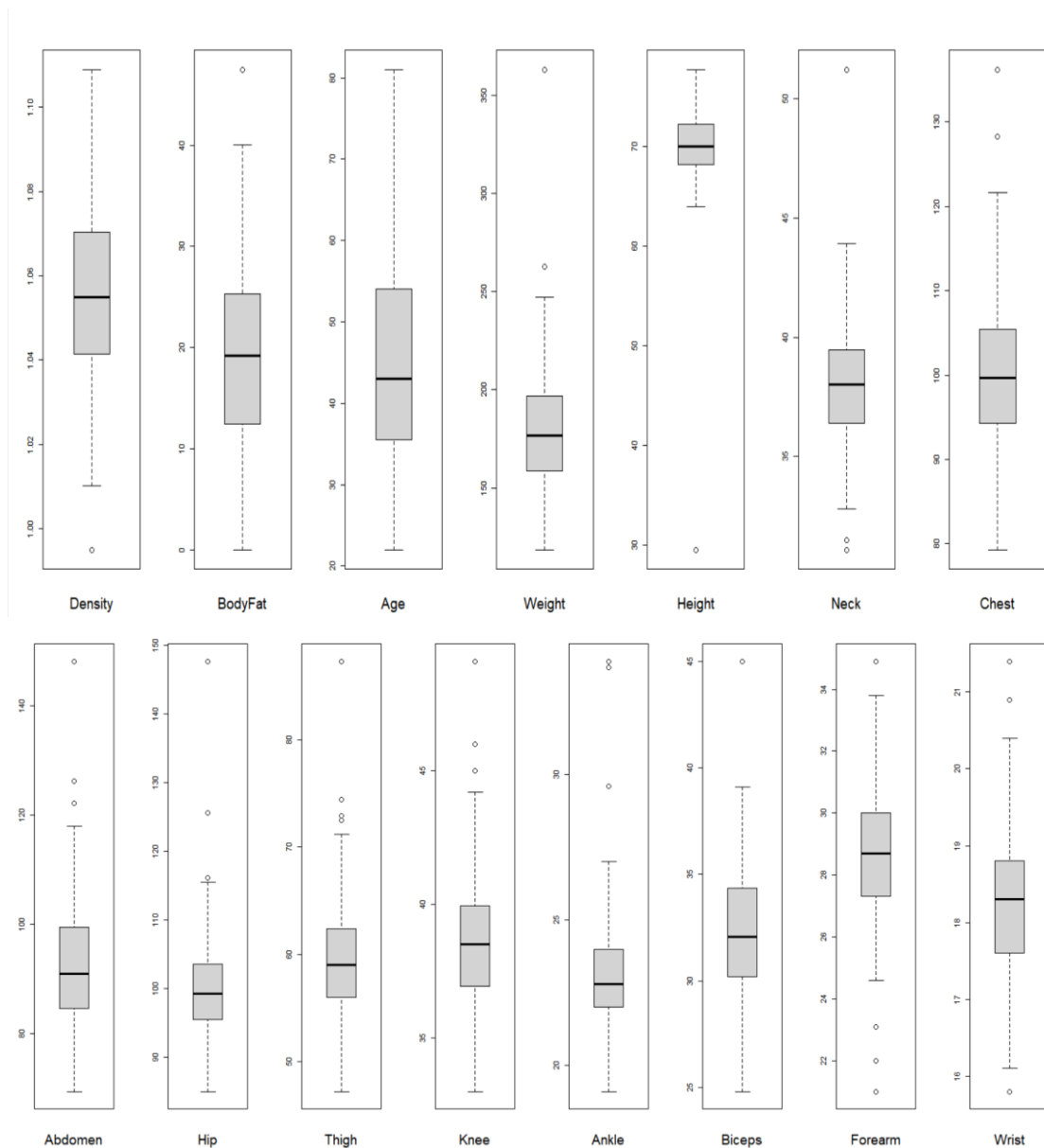
위 자료는 15개의 상당히 많은 변수가 존재한다. 15개의 변수를 모두 이용하는 것보다 자료를 잘 설명할 수 있는 최소한의 차원으로 축소하여 자료를 분석 및 해석하는 것이 더 용이하다고 판단된다. 선택된 소수의 주성분으로 신체 외부 특성을 설명함으로써 간단한 설명으로 사람들의 개별 특징을 서술할 수 있는 점도 기대된다. 주성분 분석을 통해 위 15가지의 변수를 이용하여 새로운 성분을 도출하여 해석하고, MANOVA를 통해 도출한 주성분의 점수가 연령대별로 차이가 있는지, 있다면 연령대별로 신체적 특성이 어떠한 차이가 있는지 알아보고자 한다.

## II. 본론

### - 데이터 전처리

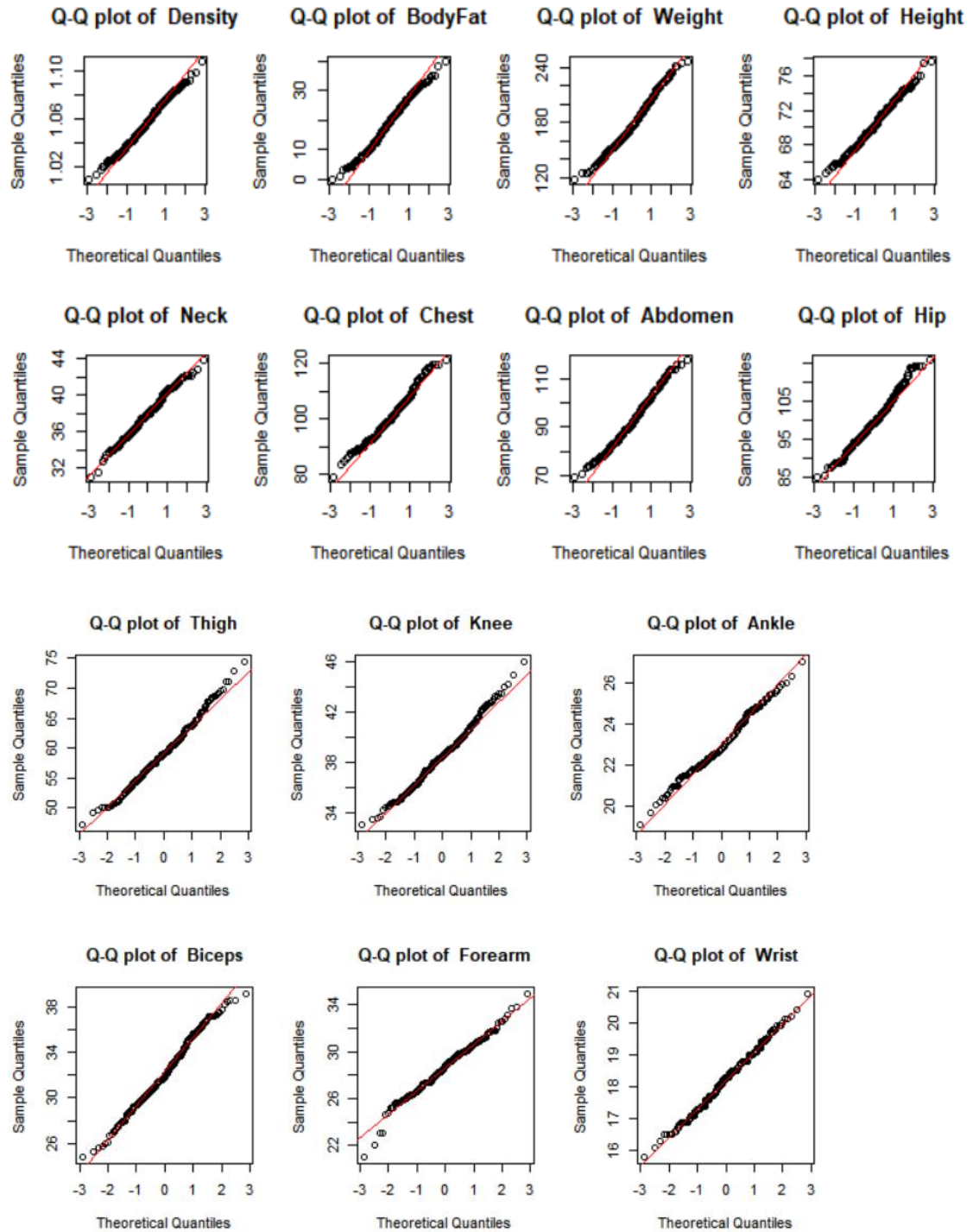
#### 1. 이상치 제거

각 변수들의 Boxplot을 통해 이상치가 존재하는지 확인하였다. Boxplot에서는 기본적으로  $(Q1-IQR*1.5, Q3+IQR*1.5)$  범위 이외의 관측치에 대해 이상치로 처리한다. 본 조는 Boxplot에서 이상치로 보여도 Boxplot에서 크게 떨어진 관측치에 대해서만 이상치로 판단하고 제거하였다.



## 2. 정규성 확인

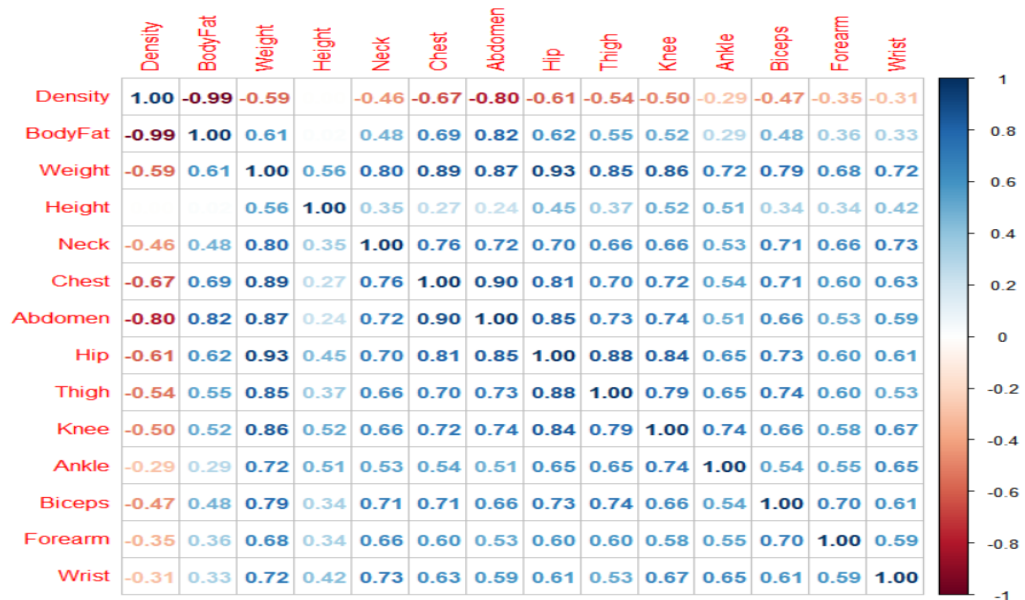
이상치를 제거한 후 자료의 정규성을 Q-Q plot을 통해 확인하였다. 나이는 이후 MANOVA에서 연령대로 범주화하여 범주형 변수로 활용할 것이므로 제외하였다.



Q-Q plot을 통해 모든 변수들의 정규성을 가정할 수 있다.

## - 주성분 분석(Principal Component Analysis)

### 1. 상관관계



Density와 Height, BodyFat과 Height를 제외한 나머지 변수들 간에는 모두 상관관계가 존재한다고 볼 수 있다.

### 2. 주성분 도출

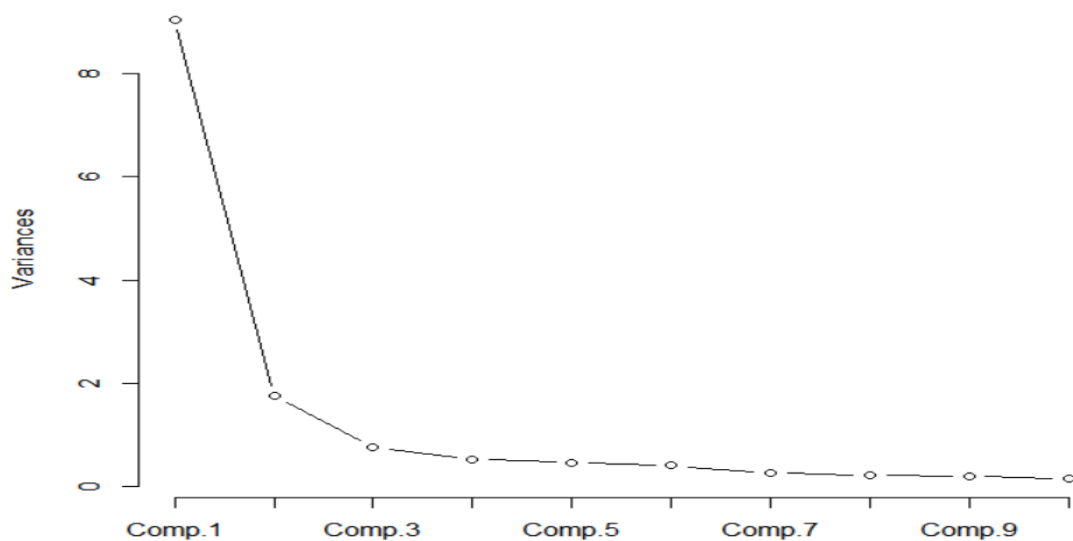
```

                Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
Standard deviation    3.008683  1.3230073  0.86870963  0.7241351  0.68146116
Proportion of Variance 0.649223  0.1255352  0.05412405  0.0376080  0.03330606
Cumulative Proportion 0.649223  0.7747582  0.82888220  0.8664902  0.89979625

                Comp.6    Comp.7    Comp.8    Comp.9    Comp.10
Standard deviation    0.6369628  0.51143800  0.4627188  0.45182883  0.3891817
Proportion of Variance 0.0290984  0.01875975  0.0153559  0.01464161  0.0108629
Cumulative Proportion 0.9288947  0.94765440  0.9630103  0.97765192  0.9885148

                Comp.11    Comp.12    Comp.13    Comp.14
Standard deviation    0.276434791  0.229598909  0.137223125  0.1103488905
Proportion of Variance 0.005480578  0.003780773  0.001350503  0.0008733271
Cumulative Proportion 0.993995397  0.997776170  0.999126673  1.0000000000

```



변수들의 단위가 다르기 때문에 자료를 표준화한 후 주성분 분석을 진행했다. 주성분의 개수를 2개로 하였을 때 설명력이 77.47%로 원자료를 설명하기에 충분하였지만, Scree plot으로 확인하였을 때는 세 번째 주성분에서 그래프가 꺾이는 것이 보이고 세 번째 주성분도 충분히 해석의 여지가 있는 것으로 판단되어 설명력 82.89%를 보이는 세 개의 주성분을 사용하였다.

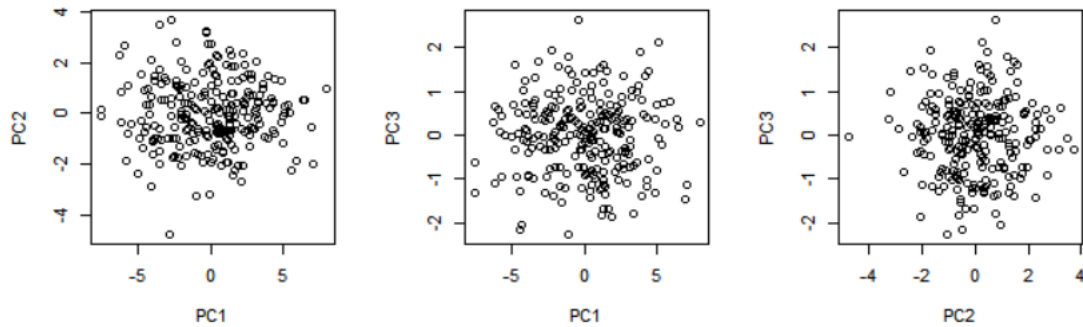
	Comp.1	Comp.2	Comp.3
Density	0.227	0.505	0.144
BodyFat	-0.232	-0.496	-0.140
Weight	-0.323		
Height	-0.152	0.473	-0.503
Neck	-0.275		0.365
Chest	-0.297	-0.134	0.102
Abdomen	-0.299	-0.243	
Hip	-0.308		-0.194
Thigh	-0.288		-0.129
Knee	-0.292	0.122	-0.245
Ankle	-0.241	0.297	-0.177
Biceps	-0.272		0.311
Forearm	-0.240	0.160	0.480
Wrist	-0.248	0.234	0.279

R의 princomp 함수의 loadings를 이용하여 3개의 주성분을 다음과 같이 정의할 수 있다.

$$\begin{aligned}
 PC1 = & 0.227 \cdot \text{Density} - 0.232 \cdot \text{BodyFat} - 0.323 \cdot \text{Weight} - 0.152 \cdot \text{Height} - 0.275 \cdot \text{Neck} \\
 & - 0.297 \cdot \text{Chest} - 0.299 \cdot \text{Abdomen} - 0.308 \cdot \text{Hip} - 0.288 \cdot \text{Thigh} - 0.292 \cdot \text{Knee} \\
 & - 0.241 \cdot \text{Ankle} - 0.272 \cdot \text{Biceps} - 0.240 \cdot \text{Forearm} - 0.248 \cdot \text{Wrist}
 \end{aligned}$$

$$\begin{aligned}
 PC2 = & 0.505 \cdot \text{Density} - 0.496 \cdot \text{BodyFat} + 0.473 \cdot \text{Height} - 0.134 \cdot \text{Chest} - 0.243 \cdot \text{Abdomen} \\
 & + 0.122 \cdot \text{Knee} + 0.297 \cdot \text{Ankle} + 0.160 \cdot \text{Forearm} + 0.234 \cdot \text{Wrist}
 \end{aligned}$$

$$\begin{aligned}
 PC3 = & 0.144*Density - 0.140*BodyFat - 0.503*Height + 0.365*Neck + 0.102*Chest \\
 & - 0.194*Hip - 0.129*Thigh - 0.245*Knee - 0.177*Ankle + 0.311*Biceps \\
 & + 0.480*Forearm + 0.279*Wrist
 \end{aligned}$$



주성분들 간의 Scatter plot을 통해 각 주성분들은 상관관계가 없는 것을 확인할 수 있다.

### 3. 주성분 의미 해석

첫 번째 주성분은 Density와는 양의 방향성, 나머지 변수들과는 음의 방향성을 가진다. 일반적으로 같은 몸무게인 두 사람 중에 지방이 더 많은 사람의 신체 둘레가 더 클 것으로 판단된다. 이 자료에 Density( $gm/cm^3$ ) 변수가 있고, Weight(pound) 변수가 있기 때문에 Weight에서 Density를 나눠 Volume을 계산하였다. 그리고 이 Volume과 첫 번째 주성분과의 상관관계를 구해본 결과 다음과 같이 매우 큰 음의 상관관계를 가졌다. 따라서 첫 번째 주성분은 (-)Volume과 다름없다는 것을 발견하였다. 따라서 첫 번째 주성분은 왜소한 정도로 볼 수 있고 값이 커질수록 왜소하다고 해석할 수 있다.

```
> cor(fat_age$volume, fat.pca$score[,1])
[1] -0.985027
```

두 번째 주성분은 Density, Height, Knee, Ankle, Forearm, Wrist와는 양의 방향성, BodyFat, Chest, Abdomen과는 음의 방향성을 가지고 특히 Density, Height와는 큰 양의 방향성, BodyFat과는 큰 음의 방향성을 가지는 것을 보아 몸에서 뼈가 차지하는 비율로 볼 수 있고 값이 커질수록 비율이 높다고 해석할 수 있다.

세 번째 주성분은 Density, Neck, Chest, Biceps, Forearm과는 양의 방향성, BodyFat, Hip, Thigh, Knee, Ankle과는 음의 방향성을 가지고 특히 키와는 큰 음의 방향성을 가지고 전반적으로 하체와는 음의 방향성, 상체와는 양의 방향성을 가진다. 일반적으로 영양소 섭취에 있어 하체가 상체보다 더 민감하고, 키도 큰 영향을 받을 것으로 예상되므로 영양소 섭취 결여 정도로 볼 수 있고 값이 커질수록 결여 정도가 크다고 해석할 수 있다.



## - MANOVA(Multivariate Analyses of Variance)

### 1. 주성분 점수에 대한 MANOVA

주성분 점수는 각 관측치가 주성분을 얼마나 가지고 있는지 알려주는 지표이다. 주성분 점수를 분석하면 위에서 해석한 주성분들이 관측치마다 어떠한 차이가 있는지 알 수 있다. MANOVA는 분석 목적에 맞게 연령대별로 주성분들이 어떻게 달라지는지 알 수 있게 해준다. 따라서 관측치들의 주성분 점수를 연령대별로 분석하는 MANOVA를 진행한 후 결과가 유의하다면 2차 분석을 통해 어떤 주성분이 연령대별로 어떠한 차이가 있는지 알아볼 것이다.

MANOVA 이전에 F 통계량이 정확하기 위해서는 자료의 정규성 가정과 등분산성 가정, 변수 간의 독립성 가정이 필요하다. 이 분석에서는 상관행렬로 구한 주성분 점수로 MANOVA를 하기 때문에 정규성과 등분산성, 독립성을 모두 만족한다. MANOVA 에서의 귀무가설은

$$H_0 : \mu_{20\text{대의 Score}} = \mu_{30\text{대의 Score}} = \mu_{40\text{대의 Score}} = \mu_{50\text{대의 Score}} = \mu_{60\text{대 이상의 Score}}$$

이다. 검정통계량 값은 데이터의 상황에 따라 4개의 검정통계량 중 선택하여 사용하는데, 표본크기가 충분하고 자료의 가정도 충족할 때 일반적으로 사용하는 Wilk의 람다를 사용하였다.

```

              Df    Wilks approx F num Df den Df      Pr(>F)
fat_age$Age    4 0.74129    6.3155     12 632.63 1.272e-10 ***
Residuals    241
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

유의수준 0.05에서 p-value가 1.272e-10로 매우 작았고, 귀무가설을 기각하여 연령대별 주성분 점수의 평균이 차이가 있음을 확인하였다.

### 2. 2차 분석

1. ANOVA		p-values			
		20s	30s	40s	50s
30s	MANOVA	.0023			
	PC1	.9011			
	PC2	.0034			
	PC3	.3137			
40s	MANOVA	.0029	.5436		
	PC1	.8591	.9054		

50s	PC2	.0002	.9246		
	PC3	.5947	.1440		
	MANOVA	<.0001	.8451	.0035	
	PC1	.6546	.6266	.4121	
	PC2	.0017	.9502	.6905	
	PC3	.0011	.4439	.0003	
Over 60	MANOVA	<.0001	<.0001	<.0001	.0002
	PC1	.2563	.3297	.2434	.1067
	PC2	<.0001	<.0001	<.0001	.0002
	PC3	.1318	.8176	<.0001	.1031

**Table1 MANOVA 사후검정표**

Table1에서 MANOVA의 p-value를 나타내는 검정의 귀무가설은

$$H_0 : \mu_{i\text{대\의 Score}} = \mu_{j\text{대\의 Score}}, i \neq j, i = 20, 30, 40, 50, \text{Over } 60, j = 30, 40, 50, \text{Over } 60$$

이고 각 주성분들의 p-value를 나타내는 검정의 귀무가설은

$$H_0 : \mu_{i\text{대\의 PC}_k \text{ Score}} = \mu_{j\text{대\의 PC}_k \text{ Score}}, i \neq j, i = 20, 30, 40, 50, \text{Over } 60, j = 30, 40, 50, \text{Over } 60,$$

$$k = 1, 2, 3$$

이다.

MANOVA에 대한 사후검정으로 연령대를 둘 씩 짝지어 사후검정을 진행하였다. Table1은 두 나이대의 MANOVA F-test의 p-value값을 나타낸 것이다. [30대와 40대] 그리고 [30대와 50대] 두 짝을 제외한 모든 짝의 p-value값이 0.05보다 작아 유의하다고 보인다. [20대와 30대]는 p-value값이 0.05보다 작긴 하지만 0.023으로 다른 연령대 짝에 비해서는 적은 차이를 보인다. 즉, [30대 및 40대], [50대의 차이]가 거의 없고 20대가 비교적 적은 차이를 내는 것으로 해석할 수 있다.

주성분 별로 비교했을 때 첫 번째 주성분은 전체 연령대를 통틀어서 p-value값이 0.05보다 크므로 차이가 나지 않는다고 볼 수 있다. 하지만 두 번째 주성분은 위의 MANOVA 결과와 같이 [30대와 40대], [30대와 50대]를 제외하고 모든 연령대 쌍에서 차이가 나기 때문에 연령대에 따라 두 번째 주성분의 차이가 크게 나는 것을 알 수 있다. MANOVA에서 연령대 쌍의 차이가 발생하는 주 원인이 두 번째 주성분임을 확인할 수 있다. 세 번째 주성분에서는 [20대와 50대], [40대와 50대], [40대와 60세 이상] 세 쌍에서는 결과가 유의하고 나머지 쌍에서는 결과가 유의하지 않다는 것을 통해 세 번째 주성분이 연령대에 따라 선형적인 관계만은 아님을 알 수 있다.

### Ⅲ. 결론

15 개의 column 으로 이루어진 체지방과 여러 신체 둘레 자료를 이용하여 주성분 분석을 진행하였고 15 개의 column 을 3 개의 주성분으로 축소할 수 있었다. 첫 번째 주성분은 왜소한 정도, 두 번째 주성분은 몸에서 뼈가 차지하는 비율, 세 번째 주성분은 영양소 섭취의 결여 정도로 해석하였다. 또, 이 주성분들의 점수를 MANOVA 를 통해 분석하였고 연령대 별로 주성분 점수의 평균이 차이가 있음을 알 수 있었다. 이후 2 차 분석을 통해 두 번째 주성분에서 주로 차이가 나는 것을 확인하였다. 이는 연령대 별로 뼈가 차지하는 비율이 차이가 있음을 의미하고, 연령대 별로 주성분 점수의 평균이 차이가 나는 주된 원인은 두 번째 주성분인 것을 알 수 있었다.

### IV. 부록

<분석에 이용한 R 코드>

```
fat=read.csv("C:/Users/yong/Desktop/다변량해석/bodyfat.csv", header=TRUE,
stringsAsFactors=TRUE)

#fat=이상치를 제거하지 않은 원자료, boxplot
par(mfrow=c(1,7))
for (i in 1:7){
    boxplot(fat[,i], xlab=names(fat)[i], cex=1.5, cex.lab=1.8)
}

par(mfrow=c(1,8))
for (i in 8:15){
    boxplot(fat[,i], xlab=names(fat)[i], cex=1.5, cex.lab=1.8)
}

#outlier row index = 31, 39, 41, 42, 86, 216 제거
fat_age=fat[-c(31,39,41,42,86,216),]
```

```

#age 제외
fat=fat_age[,-3]

#age 를 factor 로
for (i in c(20,30,40,50,60)){
    fat_age[fat_age$Age>=i & fat_age$Age<(i+10), 3]=paste(i, 's', sep='')
    if (i==60){
        fat_age[fat_age$Age>=i, 3]=paste('over', i, sep='')
    }
}
fat_age$Age=as.factor(fat_age$Age)
str(fat_age)

#histogram
par(mfrow=c(2,7))
for (i in 1:14){
    hist(fat[,i])
}

#Q-Q plot
par(mfrow=c(1,4))
par(pty='s')
for (i in 1:4){
    qqnorm(fat[,i], main=paste('Q-Q plot of ', names(fat)[i]))
    qqline(fat[,i], col='red')
}
par(mfrow=c(1,4))
par(pty='s')
for (i in 5:8){
    qqnorm(fat[,i], main=paste('Q-Q plot of ', names(fat)[i]))
    qqline(fat[,i], col='red')
}
par(mfrow=c(1,3))

```

```

par(pty='s')
for (i in 9:11){
    qqnorm(fat[,i], main=paste('Q-Q plot of ', names(fat)[i]))
    qqline(fat[,i], col='red')
}
par(mfrow=c(1,3))
par(pty='s')
for (i in 12:14){
    qqnorm(fat[,i], main=paste('Q-Q plot of ', names(fat)[i]))
    qqline(fat[,i], col='red')
}

#heatmap
library(corrplot)
corrplot(cor(fat), method='number')

#중심화 자료행렬 Y, 표준화 자료행렬 Z
X=fat
Y=X
for (i in 1:14){
    Y[,i]=X[,i]-apply(X, 2, mean)[i]
}
Z=scale(X)

#PCA
Z_pca=princomp(Z)
summary(Z_pca)
Z_pca$loadings
par(mfrow=c(1,1))
screeplot(Z_pca, type='l')

#주성분들의 독립성 확인
PC1=summary(Z_pca)$scores[,1]

```

```

PC2=summary(Z_pca)$scores[,2]
PC3=summary(Z_pca)$scores[,3]

par(mfrow=c(1,3))
par(pty='s')
plot(PC1, PC2)
plot(PC1, PC3)
plot(PC2, PC3)

#MANOVA
fat_sc <- Z_pca$scores[,1:3]
fat_m_sc <- manova(fat_sc ~ fat_age$Age)
summary(fat_m_sc)

##### Second step analysis #####

#20s and 30s
fat_age23 <- fat_age[fat_age$Age=='20s'|fat_age$Age=='30s',]
fat23 <- fat_age23[,-3]

fat23.pca <- princomp(fat23,cor=T)
fat23.sc <- fat23.pca$scores[,1:3]
fat23.sc.m <- manova(fat23.sc ~ fat_age23$Age)
summary(fat23.sc.m,test='Wilks')
summary.aov(fat23.sc.m)

#20s and 40s
fat_age24 <- fat_age[fat_age$Age=='20s'|fat_age$Age=='40s',]
fat24 <- fat_age24[,-3]

```

```

fat24.pca <- princomp(fat24,cor=T)
fat24.sc <- fat24.pca$scores[,1:3]
fat24.sc.m <- manova(fat24.sc ~ fat_age24$Age)
summary(fat24.sc.m,test='Wilks')
summary.aov(fat24.sc.m)

#20s and 50s
fat_age25 <- fat_age[fat_age$Age=='20s'|fat_age$Age=='50s',]
fat25 <- fat_age25[,-3]

fat25.pca <- princomp(fat25,cor=T)
fat25.sc <- fat25.pca$scores[,1:3]
fat25.sc.m <- manova(fat25.sc ~ fat_age25$Age)
summary(fat25.sc.m,test='Wilks')
summary.aov(fat25.sc.m)

#20s and over00
fat_age26 <- fat_age[fat_age$Age=='20s'|fat_age$Age=='over60',]
fat26 <- fat_age26[,-3]

fat26.pca <- princomp(fat26,cor=T)
fat26.sc <- fat26.pca$scores[,1:3]
fat26.sc.m <- manova(fat26.sc ~ fat_age26$Age)
summary(fat26.sc.m,test='Wilks')
summary.aov(fat26.sc.m)

#30s and 40s
fat_age34 <- fat_age[fat_age$Age=='30s'|fat_age$Age=='40s',]
fat34 <- fat_age34[,-3]

```

```

fat34.pca <- princomp(fat34,cor=T)
fat34.sc <- fat34.pca$scores[,1:3]
fat34.sc.m <- manova(fat34.sc ~ fat_age34$Age)
summary(fat34.sc.m,test='Wilks')
summary.aov(fat34.sc.m)

#30s and 50s
fat_age35 <- fat_age[fat_age$Age=='30s'|fat_age$Age=='50s',]
fat35 <- fat_age35[,-3]

fat35.pca <- princomp(fat35,cor=T)
fat35.sc <- fat35.pca$scores[,1:3]
fat35.sc.m <- manova(fat35.sc ~ fat_age35$Age)
summary(fat35.sc.m,test='Wilks')
summary.aov(fat35.sc.m)

#30s and over60
fat_age36 <- fat_age[fat_age$Age=='30s'|fat_age$Age=='over60',]
fat36 <- fat_age36[,-3]

fat36.pca <- princomp(fat36,cor=T)
fat36.sc <- fat36.pca$scores[,1:3]
fat36.sc.m <- manova(fat36.sc ~ fat_age36$Age)
summary(fat36.sc.m,test='Wilks')
summary.aov(fat36.sc.m)

#40s and 50s
fat_age45 <- fat_age[fat_age$Age=='40s'|fat_age$Age=='50s',]
fat45 <- fat_age45[,-3]

```



```

fat45.pca <- princomp(fat45,cor=T)
fat45.sc <- fat45.pca$scores[,1:3]
fat45.sc.m <- manova(fat45.sc ~ fat_age45$Age)
summary(fat45.sc.m,test='Wilks')
summary.aov(fat45.sc.m)

#40s and over60
fat_age46 <- fat_age[fat_age$Age=='40s'|fat_age$Age=='over60',]
fat46 <- fat_age46[,-3]

fat46.pca <- princomp(fat46,cor=T)
fat46.sc <- fat46.pca$scores[,1:3]
fat46.sc.m <- manova(fat46.sc ~ fat_age46$Age)
summary(fat46.sc.m,test='Wilks')
summary.aov(fat46.sc.m)

#50s and over60
fat_age56 <- fat_age[fat_age$Age=='50s'|fat_age$Age=='over60',]
fat56 <- fat_age56[,-3]

fat56.pca <- princomp(fat56,cor=T)
fat56.sc <- fat56.pca$scores[,1:3]
fat56.sc.m <- manova(fat56.sc ~ fat_age56$Age)
summary(fat56.sc.m,test='Wilks')
summary.aov(fat56.sc.m)

```