다변량통계분석 Practice 2

빅데이터경영MBA / U2016040 / 김우현

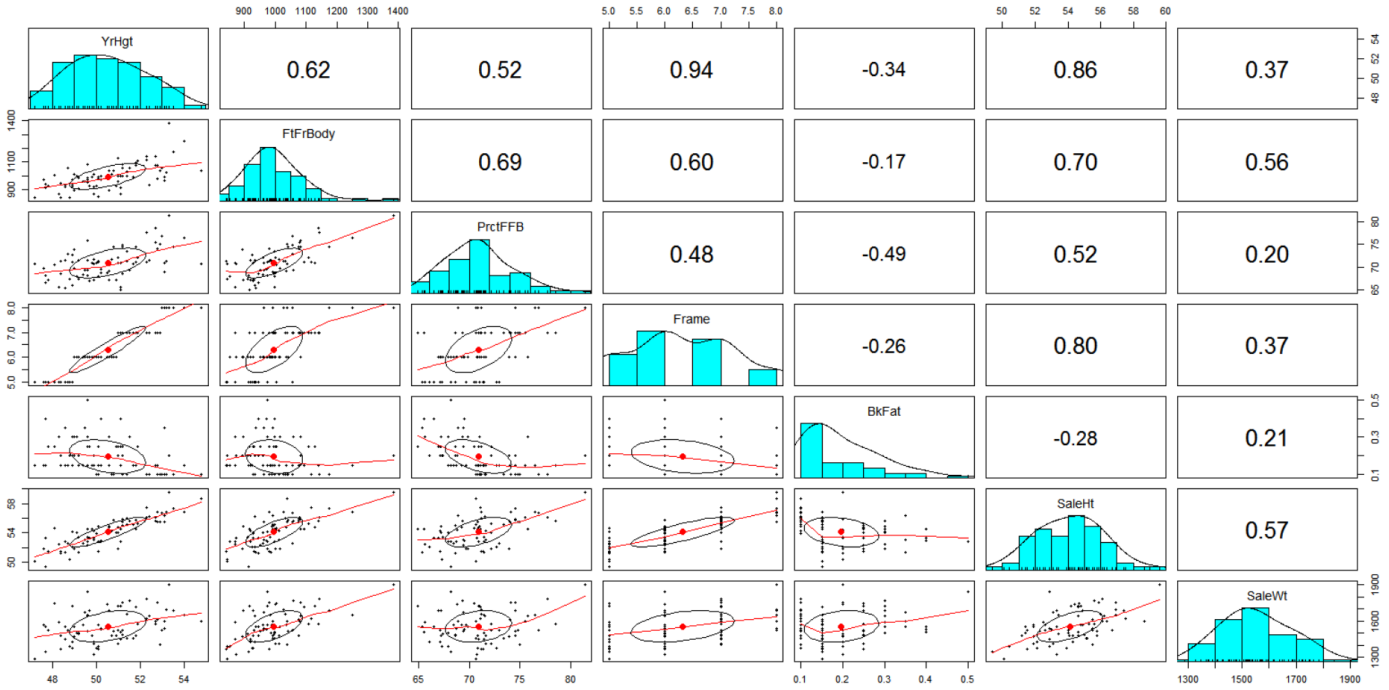
Bulls.csv는 경매시장에서 거래된 76마리의 어린(2살 이하) 황소의 특성과 거래가격(SalePr)에 관한 자료이다.

- Breed = 1 if Angus, 5 if Hereford, 8 if Simental
- FtFrBody = fat free body (pounds)
- Frame = Scale from 1(small) to 8 (large)
- SaleHt = Sale height at shoulder (inches)
- YrHgt = Yearling height at shoulder (inches)
- PrctFFB = Percent fat-free body
- BkFat = Back fat (inches)
- SaleWt = Sale weight (pounds)

SalePr와 Breed 변수를 제외한 7개의 변수를 사용해 주성분분석을 시행하여 아래의 질문에 답하시오. 공분산 행렬과 상관계수 행렬을 사용하여 각각 분석하고 비교하시오.

0. 데이터 탐색

```
data <- read.csv("bulls.csv", header = T)
head(data)
bulls <- data[, -c(1,2)]   # Breed, SalePr 변수 제거
head(bulls)
pairs.panels(bulls)
```



BkFat - 한쪽에 치우져 있어서 변환이 필요한 것처럼 보인다.

```
cor(bullsV7$YrHgt, bullsV7$BkFat)       # -0.34
cor(bullsV7$YrHgt, log(bullsV7$BkFat))  # -0.41

cor(bullsV7$SaleWt, bullsV7$BkFat)       # 0.21
cor(bullsV7$SaleWt, log(bullsV7$BkFat))  # 0.15
```

로그 변환 했으나 correlation 증가가 미미하므로 변환하지 않고 분석을 진행하는 것으로 한다.

1. 각 주성분의 표준편차와 그 주성분을 계산하는데 사용된 rotation값을 찾으시오.

(1) 공분산 행렬 이용

```
pca_cov <- prcomp(bulls)
pca_cov
```

```
Standard deviations:
[1] 143.45596038  69.81887125   2.33005787   1.82107340   0.68471173   0.27212808   0.06722679

Rotation:
                  PC1           PC2          PC3          PC4           PC5           PC6           PC7
YrHgt    -5.887328e-03 -0.0096800709 -0.286337289 -0.608787152 -0.5355689528 -0.5097273178  0.0245917521
FtFrBody -4.870470e-01 -0.8726966457  0.034277115  0.003226954 -0.0004437402 -0.0004566049 -0.0002530995
PrctFFB  -8.526499e-03 -0.0291964492 -0.904388519  0.425174911 -0.0083876301  0.0103890723  0.0142927590
Frame    -3.111988e-03 -0.0048861100 -0.133266834 -0.311194400 -0.3905733600  0.8552041268 -0.0379840767
BkFat    -6.919922e-05  0.0004925452  0.018864084  0.005278296 -0.0119061237  0.0437862261  0.9987777777
SaleHt   -9.329509e-03 -0.0085770135 -0.284214793 -0.593037047  0.7485979836  0.0823314748  0.0138200628
SaleWt   -8.732589e-01  0.4871927200 -0.004846824  0.005597435 -0.0026647979 -0.0003410092 -0.0002556156
```

➔ 주성분의 표준편차(Standard deviation)와 주성분을 계산하는데 사용된 rotation값

(2) 상관계수 행렬 이용

```
pca <- prcomp(bulls, scale = T)
pca
```

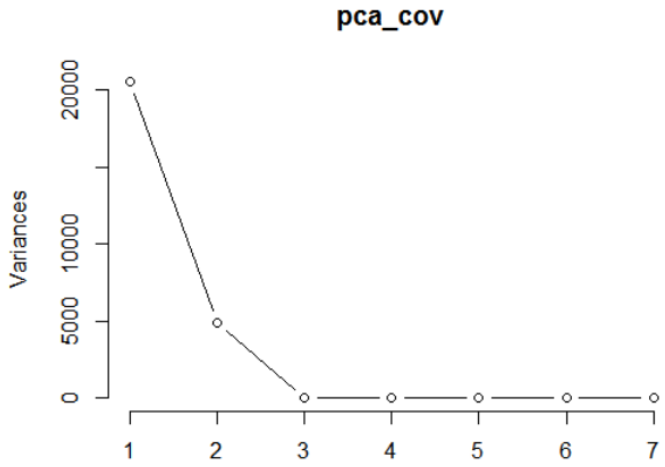
```
Standard deviations:
[1] 2.0299502 1.1563431 0.8610357 0.6491727 0.4310521 0.3827563 0.2169256

Rotation:
                PC1          PC2         PC3        PC4         PC5         PC6         PC7
YrHgt    -0.4499313 -0.042790217 -0.41570891  0.1133565 -0.06587066  0.07223418 -0.77492612
FtFrBody -0.4123256  0.129836547  0.45029241  0.2474787  0.71934339  0.17706072 -0.01776760
PrctFFB  -0.3555618 -0.315507785  0.56827313  0.3147874 -0.57936738 -0.12780009  0.00239740
Frame    -0.4339569  0.007728211 -0.45234503  0.2428179 -0.14299538  0.43414400  0.58233705
BkFat     0.1867048  0.714719363 -0.03873196  0.6181171 -0.16023789 -0.20801720 -0.04244214
SaleHt   -0.4528538  0.101315086 -0.17665043 -0.2157694  0.10953536 -0.79928778  0.23672329
SaleWt   -0.2699470  0.600514834  0.25331192 -0.5824327 -0.29054729  0.27656055 -0.04703601
```

➔ 주성분의 표준편차(Standard deviation)와 주성분을 계산하는데 사용된 rotation값

2. 적절한 주성분의 개수를 선택하고 근거를 설명하시오.

(1) 공분산 행렬 이용

```
plot(pca_cov, type = "l")
```



plot에서 제3 주성분부터 기울기가 완만하게 변하고 그 이후로는 거의 의미가 없기 때문에 제3 주성분까지 3개의 주성분을 선택한다.

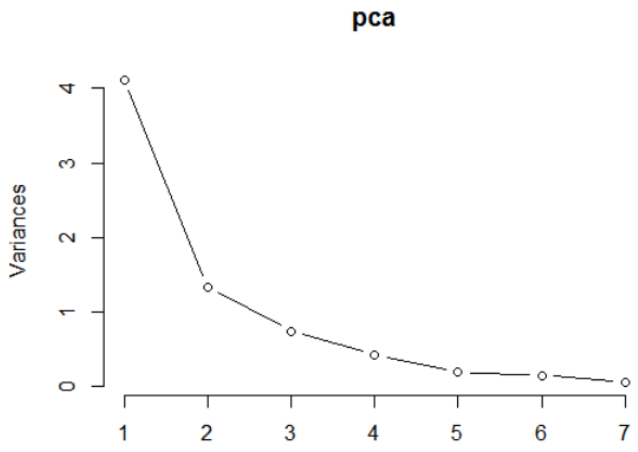
```
summary(pca_cov)
```

```
Importance of components:
                            PC1     PC2     PC3     PC4     PC5    PC6     PC7
Standard deviation     143.4560 69.8189 2.33006 1.82107 0.68471 0.2721 0.06723
Proportion of Variance   0.8082  0.1914 0.00021 0.00013 0.00002 0.0000 0.00000
Cumulative Proportion    0.8082  0.9996 0.99985 0.99998 1.00000 1.0000 1.00000
```

각 주성분의 중요도(Proportion of Variance) 누적합인 Cumulative Proportion을 보면 제1, 제2 주성분 만으로도 전체 데이터의 99.96%를 설명할 수 있다. 그러므로 2개의 주성분만 선택한다.

(2) 상관계수 행렬 이용

```
plot(pca, type = "l")
```



plot에서는 기울기가 완만하게 변하는 팔꿈치 부분을 명확하게 정하기 쉽지 않다.
제2 ~ 제5 주성분의 기울기가 비슷해 보인다.

```
summary(pca)
```

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7
Standard deviation     2.0300 1.1563 0.8610 0.6492 0.43105 0.38276 0.21693
Proportion of Variance 0.5887 0.1910 0.1059 0.0602 0.02654 0.02093 0.00672
Cumulative Proportion  0.5887 0.7797 0.8856 0.9458 0.97235 0.99328 1.00000
```

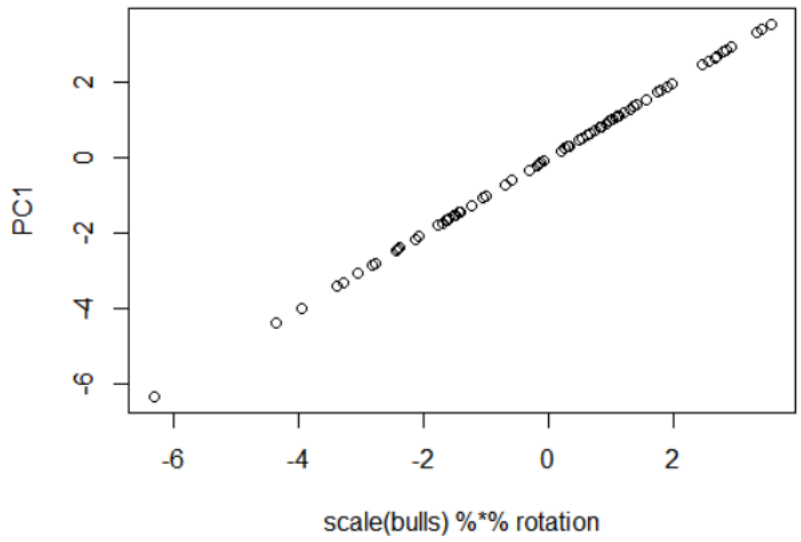
각 주성분의 중요도 누적합인 Cumulative Proportion을 보면 제3 주성분까지 선택할 경우 전체 데이터 88.56% 에 대한 설명력을 갖는다. 또한 제4 주성분의 고유값(Standard deviation)이 0.7 보다 작기 때문에 제3 주성분까지 3개의 주성분을 선택한다.

3. 각 주성분의 rotation값을 표와 그래프를 사용해 비교하고 주성분의 의미를 해석하시오.

```
a1 <- pca$rotation[,1]
a1
```

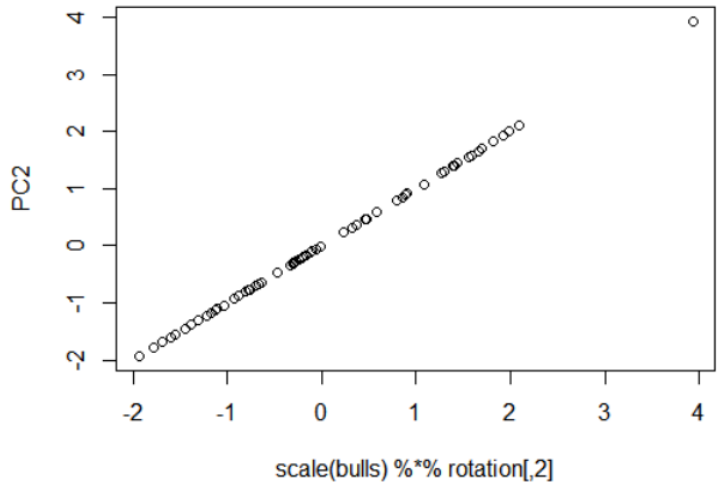
```
     YrHgt   FtFrBody    PrctFFB      Frame      BkFat     SaleHt     SaleWt
-0.4499313 -0.4123256 -0.3555618 -0.4339569  0.1867048 -0.4528538 -0.2699470
```

```
bulls_x <- scale(bullsV7) %*% a1
cor(bulls_x, pca$x[, 1])   # cor = 1

plot(bulls_x, pca$x[, 1], ylab = "PC1", xlab = "scale(bulls) %*% rotation")
```



7개의 각 변수값에 rotation 값을 행렬곱한 결과가 주성분의 값이다. 즉, 주성분의 rotation값은 각 변수에 대해 다음과 같은 설명력을 갖는다.
PC1 = (-0.4499313 * YrHgt) + (-0.4123256 * FtFrBody) + (-0.3555618 * PrctFFB) + (-0.4339569 * Frame) + (0.1867048 * BkFat) + (-0.4528538 * SaleHt) + (-0.2699470 * SaleWt)

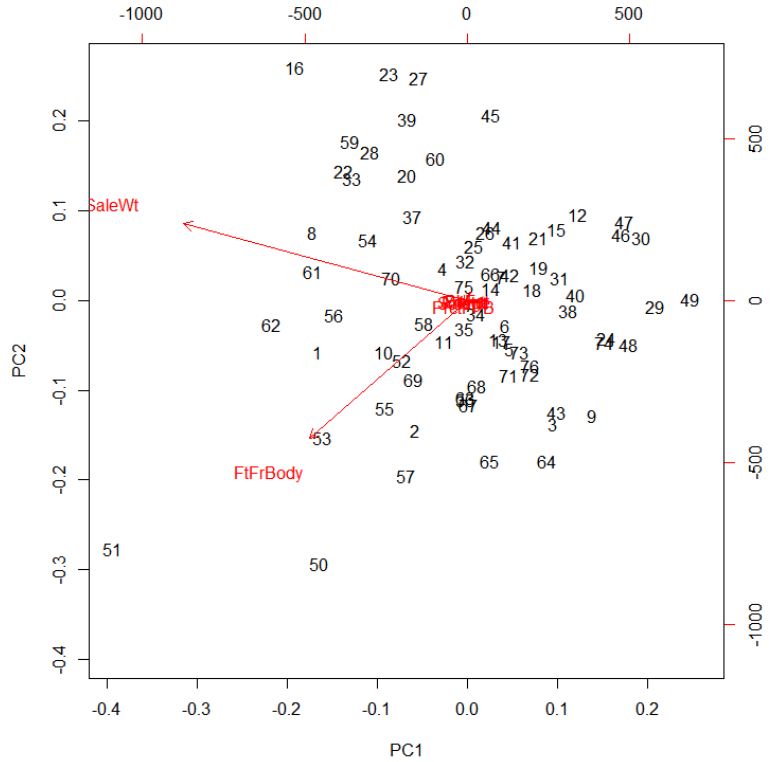
```
bulls_x2 <- scale(bullsV7) %*% pca$rotation[,2]
plot(bulls_x2, pca$x[, 2], ylab = "PC2", xlab = "scale(bulls) %*% rotation[,2]")
```



4. 행렬도를 사용해 원변수와 주성분의 관계, 원변수 간의 상관관계, 특이한 관측치의 존재 유무 등을 파악하고 설명하시오.

(1) 공분산 행렬 이용

```
biplot(pca_cov)
```



제1, 제2 주성분 만으로도 전체 데이터의 99.96%를 설명할 수 있지만 변수들의 단위가 다르기 때문에 결국 큰 값을 가진 FtFrBody와 SaleWt 두 변수가 주성분을 좌우하게 된다. 그러므로 변수간 단위가 크게 차이나는 경우 scaling이 적용되는 상관계수 행렬을 사용하는 것이 좋다.

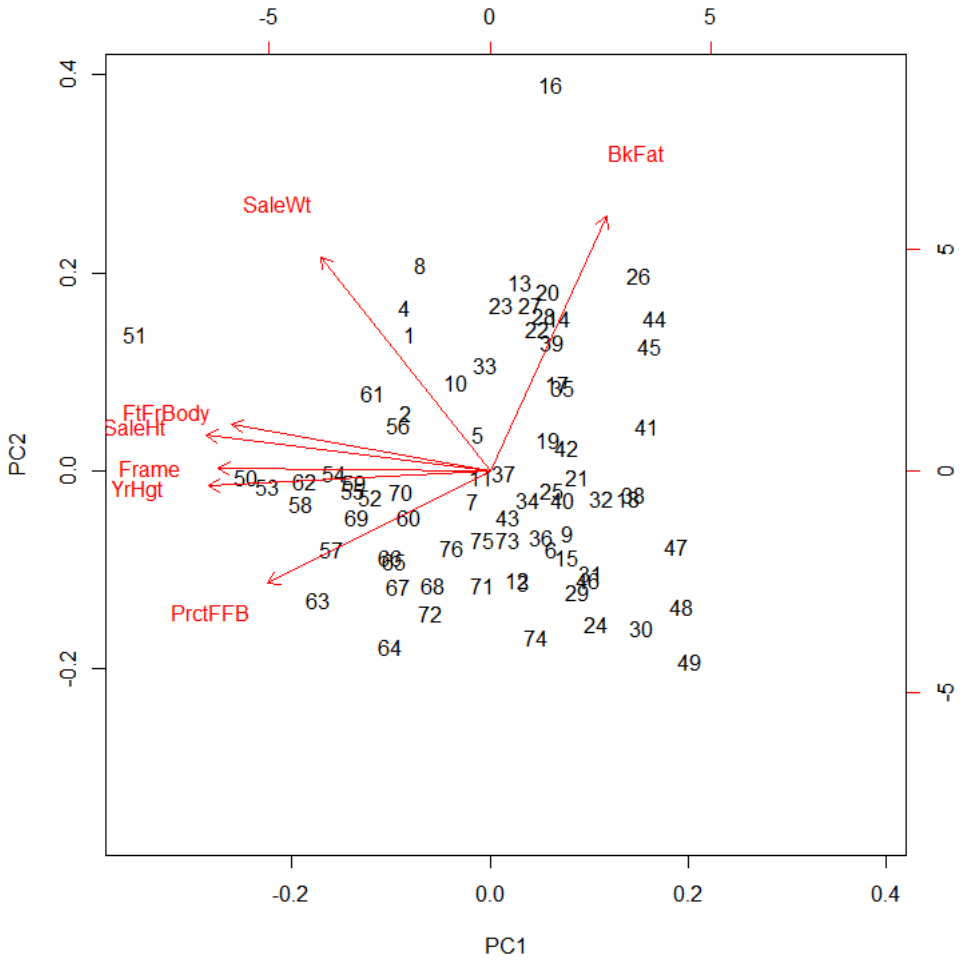
(2) 상관계수 행렬 이용

```
pca
```

```
Standard deviations:
[1] 2.0299502 1.1563431 0.8610357 0.6491727 0.4310521 0.3827563 0.2169256

Rotation:
                PC1          PC2         PC3        PC4         PC5         PC6         PC7
YrHgt     -0.4499313 -0.042790217 -0.41570891  0.1133565 -0.06587066  0.07223418 -0.77492612
FtFrBody  -0.4123256  0.129836547  0.45029241  0.2474787  0.71934339  0.17706072 -0.01776760
PrctFFB   -0.3555618 -0.315507785  0.56827313  0.3147874 -0.57936738 -0.12780009  0.00239740
Frame     -0.4339569  0.007728211 -0.45234503  0.2428179 -0.14299538  0.43414400  0.58233705
BkFat      0.1867048  0.714719363 -0.03873196  0.6181171 -0.16023789 -0.20801720 -0.04244214
SaleHt    -0.4528538  0.101315086 -0.17665043 -0.2157694  0.10953536 -0.79928778  0.23672329
SaleWt    -0.2699470  0.600514834  0.25331192 -0.5824327 -0.29054729  0.27656055 -0.04703601
```

BkFat, SaleWt를 제외한 모든 변수가 제1 주성분에 대해 비슷한 정도로 기여를 하고 있다.
BkFat, SaleWt는 제2 주성분을 특징짓는 변수이다.

```
biplot(pca)
```



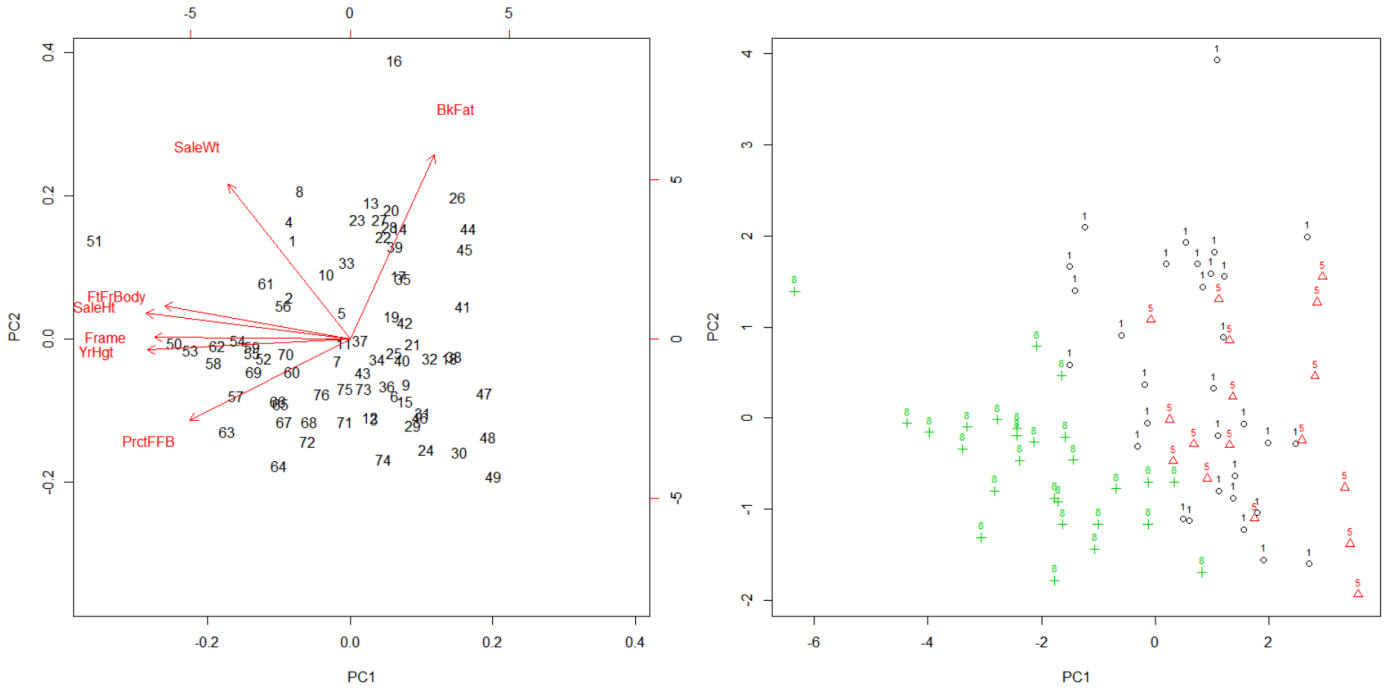
BkFat 변수는 제1 주성분의 특징과 반대되는 성향을 보인다.
즉 BkFat 값이 큰 경우 SaleHt, YrHgt, Frame 등의 값은 작은 경향을 보인다.

```
bulls[c(26, 44), ]   # BkFat 값이 큰 소
bulls[c(63, 57), ]   # PrctFFB 값이 큰 소
```

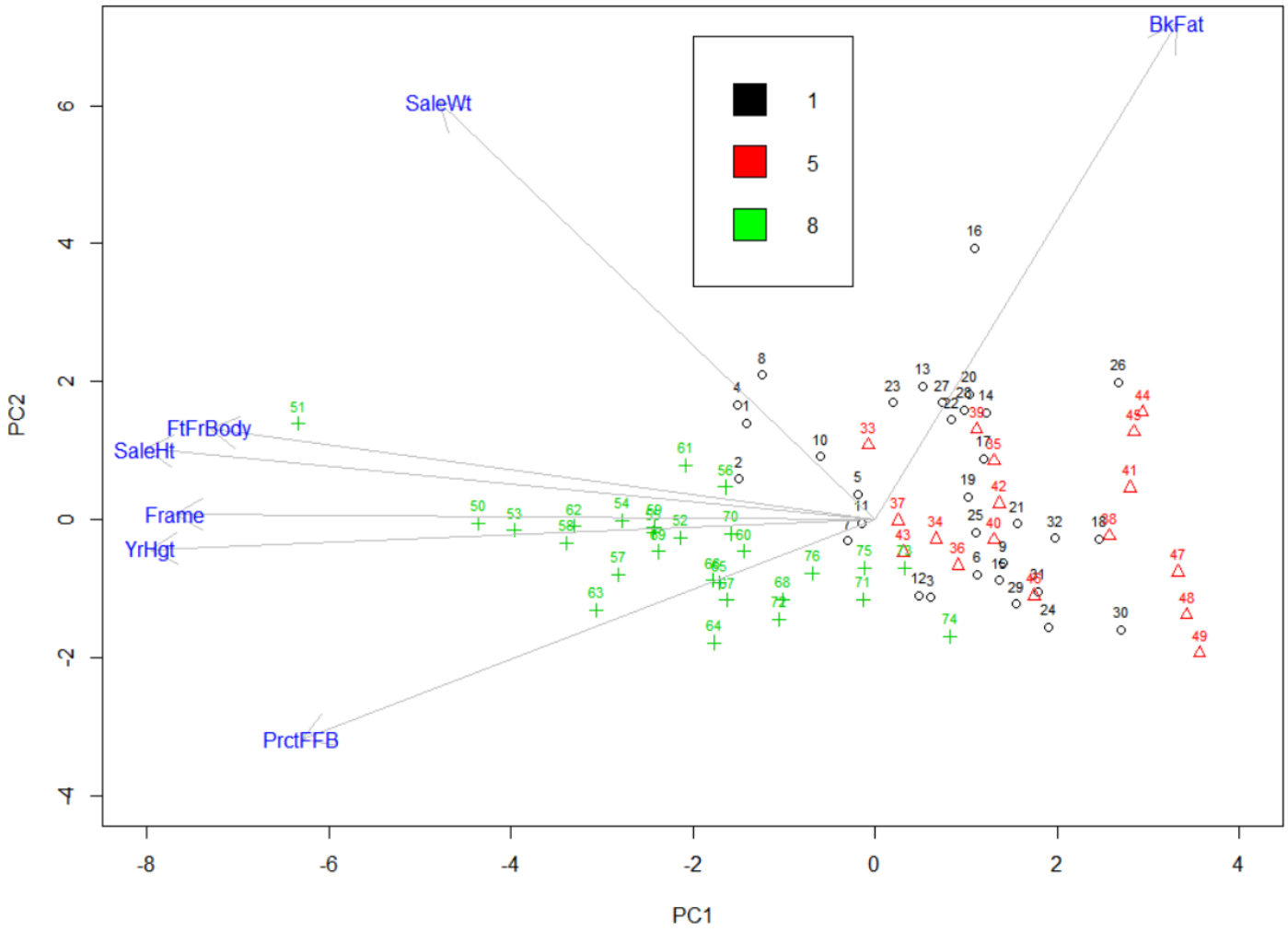
```
    Breed SalePr YrHgt FtFrBody PrctFFB Frame BkFat SaleHt SaleWt
26      1   1800  47.7      944    66.5     5  0.40   53.3   1556
44      5    975  48.6      936    65.3     5  0.35   51.4   1550
63      8   1825  53.0     1055    76.8     8  0.10   56.7   1526
57      8   1925  52.7     1141    78.5     7  0.15   55.6   1572
```

5. 첫 두개의 주성분을 사용해 산점도를 그리고 Breed를 서로 다른 색깔과 기호로 표시하시오. 주성분에 의해 다른 종의 황소를 구분할 수 있는가? 이상점이 있는가? 있다면 어떤 특성을 가진 소인가?

```
bulls$Breed <- factor(bulls$Breed)
par(mfcol = c(1,2))
biplot(pca)
plot(pca$x[,1], pca$x[,2], xlab = "PC1", ylab = "PC2",
     pch = as.numeric(bulls$Breed), col = as.numeric(bulls$Breed))
text(pca$x[,1], pca$x[,2], labels = as.character(bulls$Breed),
     cex = 0.7, pos = 3, col = as.numeric(bulls$Breed))
par(mfcol = c(1,1))
```



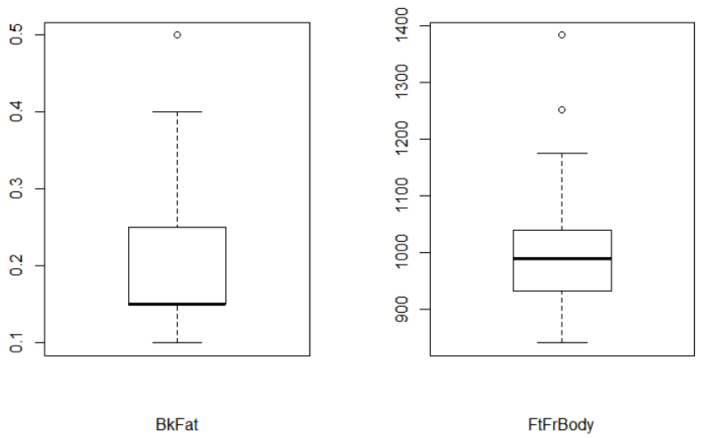

```
plot(pca$x[,1], pca$x[,2], xlab = "PC1", ylab = "PC2", xlim = c(-8,4), ylim = c(-4,7),
     pch = as.numeric(bulls$Breed), col = as.numeric(bulls$Breed))
text(pca$x[,1], pca$x[,2], labels = rownames(bulls),
     cex = 0.7, pos = 3, col = as.numeric(bulls$Breed))
legend(-2, 7, c(1,5,8), col = c("black", "red", "green"), fill = c("black", "red", "green"))
lambda <- pca$sdev * sqrt(nrow(pca$x))
Rot <- t(t(pca$rotation)*lambda)
arrows(rep(0,nrow(pca$rotation)), rep(0,nrow(pca$rotation)), Rot[,1], Rot[,2], col = "grey")
text(Rot[,1:2], rownames(Rot), col = "blue")
```



제1 주성분으로 8번종의 소는 구별해낼 수 있지만, 1번과 5번 종의 소는 비슷한 특성으로 보이며 섞여 있다. 하지만 1번종은 BkFat, SaleWt이 5번 종보다 큰 경향을 보인다.

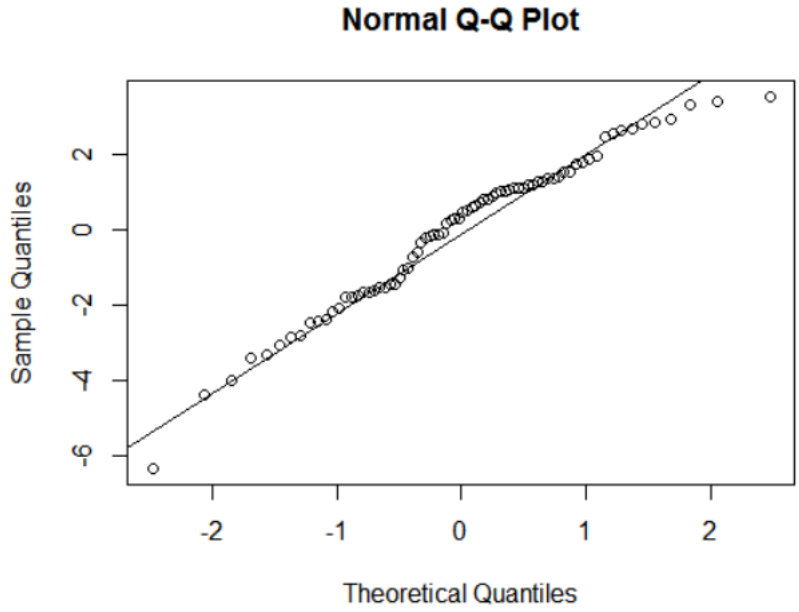
이상점(outlier)
16번 소는 BkFat에서 특이하게 큰 값을 보인다.
51번 소는 FtFrBody에서 특이하게 큰 값을 보인다.

```
boxplot(bulls$BkFat, xlab = "BkFat")
boxplot(bulls$FtFrBody, xlab = "FtFrBody")
```



6. 첫 주성분을 사용해 Q-Q plot을 그리고 해석하시오.

```
qqnorm(pca$x[,1])
qqline(pca$x[,1])
```




```
shapiro.test(pca$x[,1])
```

```
        Shapiro-Wilk normality test

data:  pca$x[, 1]
W = 0.96961, p-value = 0.06496
```

p-value > 0.05 이므로 정규분포를 따른다고 할 수 있다.
그러므로 위 주성분분석으로 도출한 주성분들은 회귀분석 등 다른 분석의 자료로 사용할 수 있다.