Wk15-2 : 주성분 회귀분석
(Principle Component Regression)

1. 주성분회귀 (Principle Component Regression)

- 독립변수들의 차원을 줄이기 위해 사용가능, 주성분을 이용하여 타겟변수(Y)의 설명력(예측력)을 높일수 있다.

$$Y = b_0 + b_1 PC_1 + b_2 PC_2$$

- 독립변수들의 전체분산을 가장 잘 설명해주는 component를 사용하여 독립변수들간 다중공선성 문제를 해결할 수 있다.
- 주요 component score들이 Y의 예측력을 보장하는 것은 아니다. 주요 component score는 X의 분산을 가장 잘 설명하는 방향의 축을 기준으로 변환된 것이기 때문에 Y와의 관계에 있어서는 상관성이 없을수도 있다.

1. 주성분회귀 (Principle Component Regression)

- wine데이터 (9개의 독립변수, 타겟변수는 Aroma rating)

9개의 독립변수

타겟변수(y) : Aroma

	Mo	Ba	Cr	Sr	Pb	B	Mg	Ca	K	Aroma
2	0.044	0.387	0.029	1.23	0.561	2.63	128	80.5	1130	3.3
3	0.16	0.312	0.038	0.975	0.697	6.21	193	75	1010	4.4
4	0.146	0.308	0.035	1.14	0.73	3.05	127	91	1160	3.9
5	0.191	0.165	0.036	0.927	0.796	2.57	112	93.6	924	3.9
6	0.363	0.38	0.059	1.13	1.73	3.07	138	84.6	1090	5.6
7	0.106	0.275	0.019	1.05	0.491	6.56	172	112	1290	4.6
8	0.479	0.164	0.062	0.823	2.06	4.57	179	122	1170	4.8
9	0.234	0.271	0.044	0.963	1.09	3.18	145	91.9	1020	5.3
10	0.058	0.225	0.022	1.13	0.048	6.13	113	70.2	1240	4.3

1. 주성분회귀 (Principle Component Regression)

- wine 데이터 불러들이기, 독립변수들간 상관계수

```
# lec15_2_pcr.r
# Multivariate analysis
# Principle Component regression

# set working directory
setwd("D:/tempstore/moocr/wk15")

# wine data
wine<-read.csv(file="wine_aroma.csv")
attach(wine)
head(wine)
```

wine 데이터 불러들이기

head(wine)-첫번째부터 6번째 데이터 보여줌

```
# Check correlation
cor(wine[1:9])
```

독립변수간 상관관계 확인

0.95, 0.82등 일부 높은 상관계수

```
> cor(wine[1:9])
           Mo          Ba          Cr          Sr          Pb          B
Mo  1.000000000 -0.005845794 0.43202327 -0.04506390  0.95042130 -0.10139057
Ba -0.005845794  1.000000000 0.03307179  0.82068003  0.09127673 -0.25386503
Cr  0.432023266  0.033071792 1.00000000  0.24604133  0.35372849  0.08280968
Sr -0.045063901  0.820680029 0.24604133  1.00000000 -0.03084108 -0.32126815
Pb  0.950421296  0.091276726 0.35372849 -0.03084108  1.00000000 -0.19147512
B  -0.101390567 -0.253865031 0.08280968 -0.32126815 -0.19147512  1.00000000
```

1. 주성분회귀 (Principle Component Regression)

• 주성분분석을 위한 함수 : prcomp(독립변수들, center= , scale=)

```
# 1. PCA(center=T->mean=0, scale.=T->variance=1)
wi.pca<-prcomp(wine[1:9],center=T,scale.=F)
# wi.pca<-prcomp(wine[1:9],center=T,scale.=T)
wi.pca
summary(wi.pca)
```

- 옵션을 주지않으면 center=T, scale=F
- center=T, scale=F는 mean-centering만 한다음 component를 뽑음

```
> wi.pca<-prcomp(wine[1:9],center=T,scale.=F)
> # wi.pca<-prcomp(wine[1:9],center=T,scale.=T)
> wi.pca
Standard deviations (1, .., p=9):
[1] 191.34178051  27.33235332  15.40582103   1.57383189   0.41900698
[6]   0.17041133   0.05942324   0.02365329   0.01089562

Rotation (n x k) = (9 x 9):
              PC1           PC2           PC3           PC4           PC5
Mo  4.396532e-05  1.531130e-03 -0.0016925110  6.688447e-05 -0.188670999
Ba -2.644028e-04  1.588626e-03  0.0016225184  2.310870e-02  0.006670829
Cr -4.594881e-05  3.640528e-05  0.0002075367  5.172124e-04 -0.022994509
Sr -1.386527e-03  7.314677e-03  0.0074873569  1.258356e-01  0.183855603
Pb  2.621301e-04  6.962105e-03 -0.0076524788  2.814073e-02 -0.964311791
B  -2.516393e-03 -1.071784e-02  0.0005113384 -9.913097e-01 -0.003970591
Mg -7.437057e-02  9.302614e-01  0.3589981755 -1.087874e-02  0.001918085
Ca -4.369679e-02  3.565893e-01 -0.9331669873 -3.859202e-03  0.010450686
K  -9.962686e-01 -8.506499e-02  0.0141159901  3.311344e-03 -0.001110159
```

PC1 = 0.0000439*Mo -0.00026* Ba - 0.000046*Cr +....

1. 주성분회귀 (Principle Component Regression)

• 전체분산 중 각 주성분의 설명하는 비율

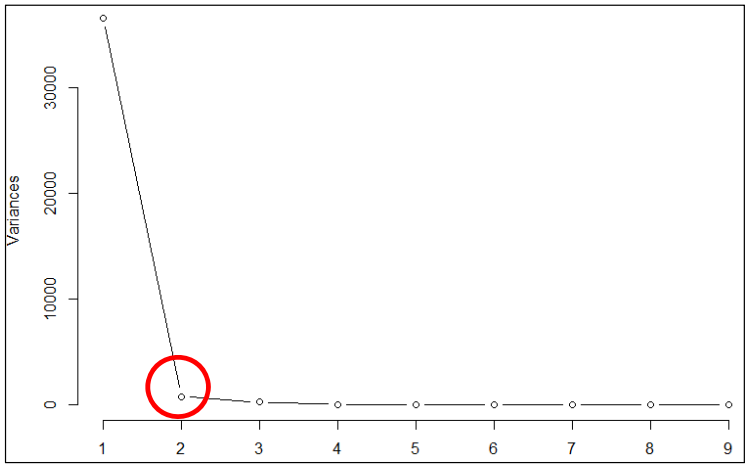
```
> summary(wi.pca)
Importance of components:
                          PC1      PC2      PC3     PC4   PC5    PC6     PC7
Standard deviation     191.3418 27.33235 15.40582 1.57383 0.419 0.1704 0.05942
Proportion of Variance   0.9738  0.01987  0.00631 0.00007 0.000 0.0000 0.00000
Cumulative Proportion    0.9738  0.99362  0.99993 0.99999 1.000 1.0000 1.00000
                           PC8     PC9
Standard deviation     0.02365 0.0109
Proportion of Variance 0.00000 0.0000
Cumulative Proportion  1.00000 1.0000
```

- PC1은 전체분산의 97.38%를 설명, 1개의 변수만으로도 독립변수 전체분산을 거의 설명
- PC2은 전체분산의 1.98%를 설명
- 누적설명비율을 보면 PC1와 PC2, 두개의 성분으로 전체분산의 99.36%를 설명

1. 주성분회귀 (Principle Component Regression)

• 최적 주성분 수는? – scree plot을 그려보고 급격히 떨어지기 전까지의 PC를 선택

```
# 2.scree plot : to choose the number of components
plot(wi.pca,type="l")
```



- 2rd PC에서 설명력이 급격하게 떨어짐을 볼 수 있음

=> 이 경우 PC1 한 개만 사용해도 된다는 의미

1. 주성분회귀 (Principle Component Regression)

• PC계산= X_data(n*p) %*% PCA_weight(p*p)

```
# 3. calculate component=x_data%*% PCA weight
PRC<-as.matrix(wine[,1:9])%*%wi.pca$rotation
head(PRC)
```

PRC는 n*p행렬, 여기서는 37*9
head(PRC)는 첫번째 6줄을 보여줌

```
> PRC<-as.matrix(wine[,1:9])%*%wi.pca$rotation
> head(PRC)
           PC1       PC2       PC3        PC4        PC5       PC6
[1,] -1138.8289  51.64087 -13.21029 -0.3889438 -0.4993510 0.4769412
[2,] -1023.8790 120.31518  13.56166 -5.0510679 -0.5137726 0.7069155
[3,] -1169.1022  51.89889 -22.94611 -0.7439815 -0.6259345 0.4848385
[4,]  -932.9793  58.95133 -34.09134 -0.9247297 -0.4759034 0.3602776
[5,] -1099.9016  65.81126 -14.02094 -1.0619959 -1.6012517 0.3328149
[6,] -1302.8902  90.15052 -24.54966 -4.3824147 -0.2567956 0.6480738
```

PC1 = 0.0000439*Mo -0.00026* Ba - 0.000046*Cr +....

2. 주성분을 이용한 회귀모형

- wine data => wine.pc data구성

```
# 4. Principal component regression
# make data with components
wine.pc<-cbind(as.data.frame(PRC),Aroma)
head(wine.pc)
```

wine 데이터

```
> head(wine)
     Mo    Ba    Cr    Sr    Pb    B  Mg    Ca    K Aroma
1 0.044 0.387 0.029 1.230 0.561 2.63 128  80.5 1130   3.3
2 0.160 0.312 0.038 0.975 0.697 6.21 193  75.0 1010   4.4
3 0.146 0.308 0.035 1.140 0.730 3.05 127  91.0 1160   3.9
4 0.191 0.165 0.036 0.927 0.796 2.57 112  93.6  924   3.9
5 0.363 0.380 0.059 1.130 1.730 3.07 138  84.6 1090   5.6
6 0.106 0.275 0.019 1.050 0.491 6.56 172 112.0 1290   4.6
```

iris.pc 데이터

```
> head(wine.pc)
          PC1       PC2       PC3        PC4        PC5
1 -1138.8289  51.64087 -13.21029 -0.3889438 -0.4993510
2 -1023.8790 120.31518  13.56166 -5.0510679 -0.5137726
3 -1169.1022  51.89889 -22.94611 -0.7439815 -0.6259345
4  -932.9793  58.95133 -34.09134 -0.9247297 -0.4759034
5 -1099.9016  65.81126 -14.02094 -1.0619959 -1.6012517
6 -1302.8902  90.15052 -24.54966 -4.3824147 -0.2567956
```

PC1 = 0.0000439*Mo -0.00026* Ba - 0.000046*Cr +....

2. 주성분을 이용한 회귀모형

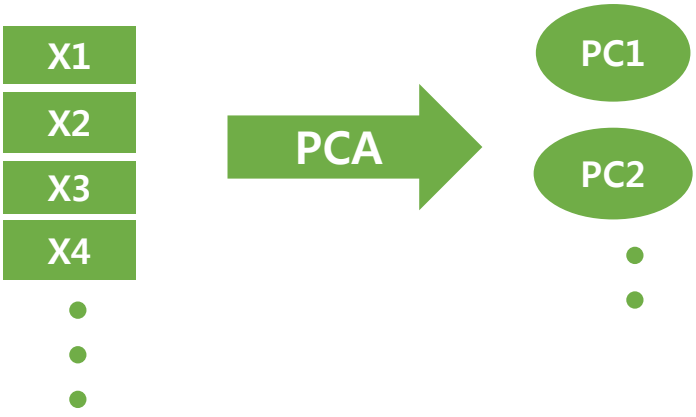
- 다중회귀모형과 주성분회귀분석

다중회귀모형

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_k X_k$$

주성분회귀모형

$$Y = \beta_0 + \beta_1 PC_1 + \beta_2 PC_2 + ..$$



f(PC1, PC2, ...)=y

타겟값(Y)를 가장 잘 예측하는 선형모형

2. 주성분을 이용한 회귀모형

• 주성분을 이용한 회귀분석모형 1 (wine data : PC1-PC4포함)

```
# regression(PC1-PC4)
fit1<-lm(Aroma~PC1+PC2+PC3+PC4, data=wine.pc)
fit1
summary(fit1)
```

```
> summary(fit1)

Call:
lm(formula = Aroma ~ PC1 + PC2 + PC3 + PC4, data = wine.pc)

Residuals:
     Min       1Q   Median       3Q      Max
-1.37626 -0.66068  0.00352  0.48748  1.35150

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.7499387  0.8204063   9.446 8.96e-11 ***
PC1          0.0025532  0.0006629   3.852 0.000530 ***
PC2         -0.0147475  0.0046404  -3.178 0.003279 **
PC3          0.0031120  0.0082328   0.378 0.707924
PC4         -0.3022774  0.0805886  -3.751 0.000701 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.761 on 32 degrees of freedom
Multiple R-squared:  0.5502,    Adjusted R-squared:  0.494
F-statistic: 9.787 on 4 and 32 DF,  p-value: 2.749e-05
```

PC1-PC4 다중회귀모형 수행
(R^2=.494)

2. 주성분을 이용한 회귀모형

• 주성분을 이용한 회귀분석모형 2 (wine data : PC1-PC9 포함)

```
# regression(PC1-PC9)
fit2<-lm(Aroma~., data=wine.pc)
fit2
summary(fit2)
```

```
Call:
lm(formula = Aroma ~ ., data = wine.pc)

Residuals:
     Min       1Q   Median       3Q      Max
-0.86916 -0.32577 -0.02585  0.25665  0.92525

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.738e+00  7.174e-01   7.998 1.35e-08 ***
PC1          2.553e-03  4.741e-04   5.386 1.08e-05 ***
PC2         -1.475e-02  3.319e-03  -4.444 0.000136 ***
PC3          3.112e-03  5.888e-03   0.529 0.601455
PC4         -3.023e-01  5.764e-02  -5.244 1.58e-05 ***
PC5         -9.251e-01  2.165e-01  -4.273 0.000214 ***
PC6          1.695e+00  5.323e-01   3.184 0.003639 **
PC7         -1.002e+00  1.527e+00  -0.656 0.517085
PC8         -7.910e+00  3.835e+00  -2.062 0.048902 *
PC9         -1.309e+01  8.326e+00  -1.573 0.127409
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5443 on 27 degrees of freedom
Multiple R-squared:  0.8059,    Adjusted R-squared:  0.7412
```

PC1-PC9 다중회귀모형 수행
(R^2=.741)

2. 주성분을 이용한 회귀모형

- 일반 회귀분석모형 (wine data : raw data)

```
# Multiple regression with the raw data
fit3<-lm(Aroma ~., data=wine)
summary(fit3)
```

```
Call:
lm(formula = Aroma ~ ., data = wine)

Residuals:
     Min       1Q   Median       3Q      Max
-0.86916 -0.32577 -0.02585  0.25665  0.92525

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.737960   0.717394   7.998 1.35e-08 ***
Mo           9.207592   3.807472   2.418 0.022612 *
Ba           1.541877   1.992476   0.774 0.445743
Cr         -12.039330   8.196114  -1.469 0.153414
Sr          -2.397588   0.652929  -3.672 0.001047 **
Pb          -1.001847   0.779480  -1.285 0.209612
B            0.002493   0.094840   0.026 0.979218
Mg           0.004604   0.006227   0.739 0.466116
Ca          -0.029251   0.007610  -3.844 0.000668 ***
K            0.001996   0.001122   1.778 0.086641 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5443 on 27 degrees of freedom
Multiple R-squared:  0.8059,    Adjusted R-squared:  0.7412
```

9개 독립변수 다중회귀모형 수행 (R^2=.741)

2. 주성분을 이용한 회귀모형

- 잔차에 대한 가정 확인

```
# residual diagnostic plot
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(fit3)
```