주성분 분석을 위한 보충 R 코드와 결과

이재용, 임요한

서울대학교
통계학과

September 18, 2017

Contents

1 준비

1.1 전역옵션들

```
opts_chunk$set(eval=TRUE, cache=TRUE, fig.width=7, fig.height=4)
```

1.2 패키지 로딩

```
library(ISLR)
```

2 주성분분석

```
states=row.names(USArrests)
states
```

```
##  [1] "Alabama"        "Alaska"         "Arizona"        "Arkansas"
##  [5] "California"     "Colorado"       "Connecticut"    "Delaware"
##  [9] "Florida"        "Georgia"        "Hawaii"         "Idaho"
## [13] "Illinois"       "Indiana"        "Iowa"           "Kansas"
## [17] "Kentucky"       "Louisiana"      "Maine"          "Maryland"
## [21] "Massachusetts"  "Michigan"       "Minnesota"      "Mississippi"
## [25] "Missouri"       "Montana"        "Nebraska"       "Nevada"
## [29] "New Hampshire"  "New Jersey"     "New Mexico"     "New York"
## [33] "North Carolina" "North Dakota"   "Ohio"           "Oklahoma"
## [37] "Oregon"         "Pennsylvania"   "Rhode Island"   "South Carolina"
## [41] "South Dakota"   "Tennessee"      "Texas"          "Utah"
## [45] "Vermont"        "Virginia"       "Washington"     "West Virginia"
## [49] "Wisconsin"      "Wyoming"
```

```
str(USArrests)
```

```
## 'data.frame': 50 obs. of  4 variables:
##  $ Murder  : num  13.2 10 8.1 8.8 9 7.9 3.3 5.9 15.4 17.4 ...
##  $ Assault : int  236 263 294 190 276 204 110 238 335 211 ...
##  $ UrbanPop: int  58 48 80 50 91 78 77 72 80 60 ...
##  $ Rape    : num  21.2 44.5 31 19.5 40.6 38.7 11.1 15.8 31.9 25.8 ...
```

```
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

자료의 형태를 본다. 네 개의 변수, 50개의 관측치로 구성된 자료이다.

```
apply(USArrests, 2, mean)
```

```
##   Murder  Assault UrbanPop     Rape
##    7.788  170.760   65.540   21.232
```

```
apply(USArrests, 2, sd)
```

```
##    Murder   Assault  UrbanPop      Rape
##  4.355510 83.337661 14.474763  9.366385
```

```
summary(USArrests)
```

```
##       Murder          Assault         UrbanPop          Rape
##  Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30
##  1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07
##  Median : 7.250   Median :159.0   Median :66.00   Median :20.10
##  Mean   : 7.788   Mean   :170.8   Mean   :65.54   Mean   :21.23
##  3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:77.75   3rd Qu.:26.18
##  Max.   :17.400   Max.   :337.0   Max.   :91.00   Max.   :46.00
```

각 변수들의 평균과 분산은 매우 다르다. 주성분분석을 할 때, 척도화(scaling)를 하는 것이 필요하다.

```
pr.out=prcomp(USArrests, scale=TRUE)
str(pr.out)
```

```
## List of 5
##  $ sdev    : num [1:4] 1.575 0.995 0.597 0.416
##  $ rotation: num [1:4, 1:4] -0.536 -0.583 -0.278 -0.543 0.418 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:4] "Murder" "Assault" "UrbanPop" "Rape"
##   .. ..$ : chr [1:4] "PC1" "PC2" "PC3" "PC4"
##  $ center  : Named num [1:4] 7.79 170.76 65.54 21.23
##   ..- attr(*, "names")= chr [1:4] "Murder" "Assault" "UrbanPop" "Rape"
##  $ scale   : Named num [1:4] 4.36 83.34 14.47 9.37
##   ..- attr(*, "names")= chr [1:4] "Murder" "Assault" "UrbanPop" "Rape"
##  $ x       : num [1:50, 1:4] -0.976 -1.931 -1.745 0.14 -2.499 ...
##   ..- attr(*, "dimnames")=List of 2
```

```
##   .. ..$ : chr [1:50] "Alabama" "Alaska" "Arizona" "Arkansas" ...
##   .. ..$ : chr [1:4] "PC1" "PC2" "PC3" "PC4"
##  - attr(*, "class")= chr "prcomp"
```

척도화(scale=TRUE)를 해서 주성분분석을 하였다. 주성분분석의 결과는 5개의 리스트(sdev, rotation, center, scale, x)로 이루어져있다.

```
pr.out$center
```

```
##   Murder  Assault UrbanPop     Rape
##    7.788  170.760   65.540   21.232
```

```
pr.out$scale
```

```
##    Murder   Assault  UrbanPop      Rape
##  4.355510 83.337661 14.474763  9.366385
```

5개의 리스트 중 center와 scale은 주성분분석을 수행하기 전 각 변수들의 평균과 표준편차를 의미한다.

```
pr.out$rotation
```

```
##                  PC1        PC2        PC3         PC4
## Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
## Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
## UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
## Rape     -0.5434321 -0.1673186  0.8177779  0.08902432
```

rotation은 적재 벡터를 포함하고 있다. 즉,

$$
\begin{aligned}
pc_1 &= -0.535 \times Murder - 0.583 \times Assault - 0.278 \times UrbanPop - 0.543 \times Rape \\
pc_2 &= 0.418 \times Murder + 0.187 \times Assault - 0.872 \times UrbanPop - 0.167 \times Rape
\end{aligned}
$$

와 같이 나타낼 수 있따.

```
dim(pr.out$x)
```

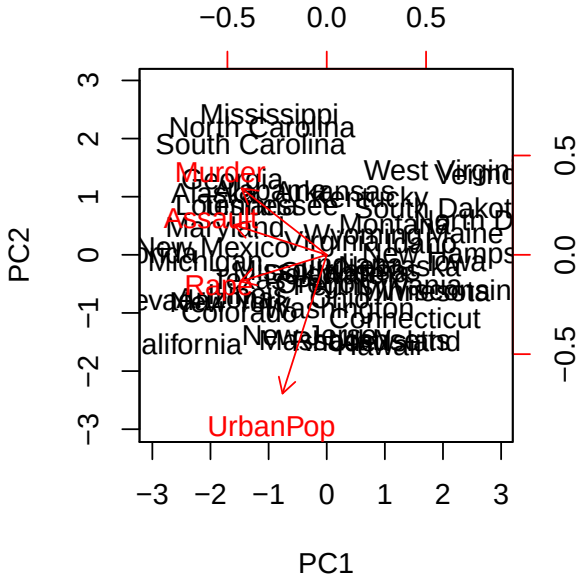
```
## [1] 50  4
```

```
str(pr.out$x)
```

```
##  num [1:50, 1:4] -0.976 -1.931 -1.745 0.14 -2.499 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:50] "Alabama" "Alaska" "Arizona" "Arkansas" ...
##   ..$ : chr [1:4] "PC1" "PC2" "PC3" "PC4"
```

x는 주성분을 포함하고 있다. 즉, x[,1]은 첫번째 주성분, x[2]는 두번째 주성분이다. 각 관측치마다 주성분의 값을 계산하여서 x의 행은 관측치의 개수와 같고 열은 주성분의 개수와 같다.

```
biplot(pr.out, scale=0)
```

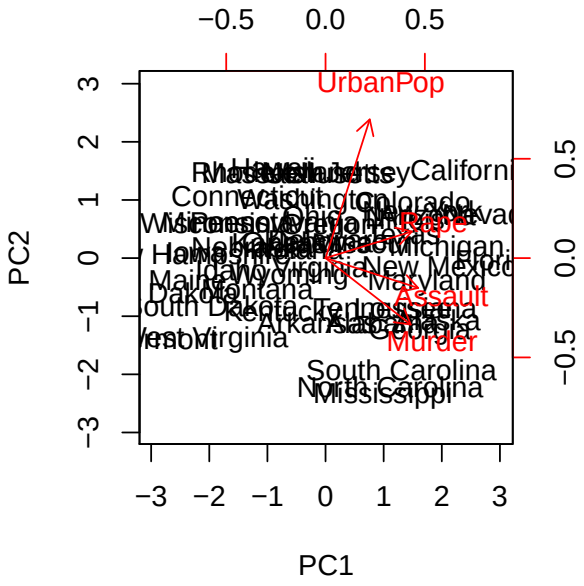


쌍도를 그린 그림이다. 이 그림은

```
plot(pr.out$x[,1], pr.out$x[,2])
```

를 그린 것과 같다. 두번째와 세번째 주성분의 그림을 그리고 싶으면 옵션 choices = c(2,3)을 쓰면 된다. 변수는 λ^{scale}와 같이 관측치는 $\lambda^{1-scale}$로 표시된다. scale=0는 있는 그대로 척도를 맞추는 것이다. λ는 주성분분석의 특이값(singular value)를 의미한다.

```
pr.out$rotation=-pr.out$rotation
pr.out$x=-pr.out$x
biplot(pr.out, scale=0)
```

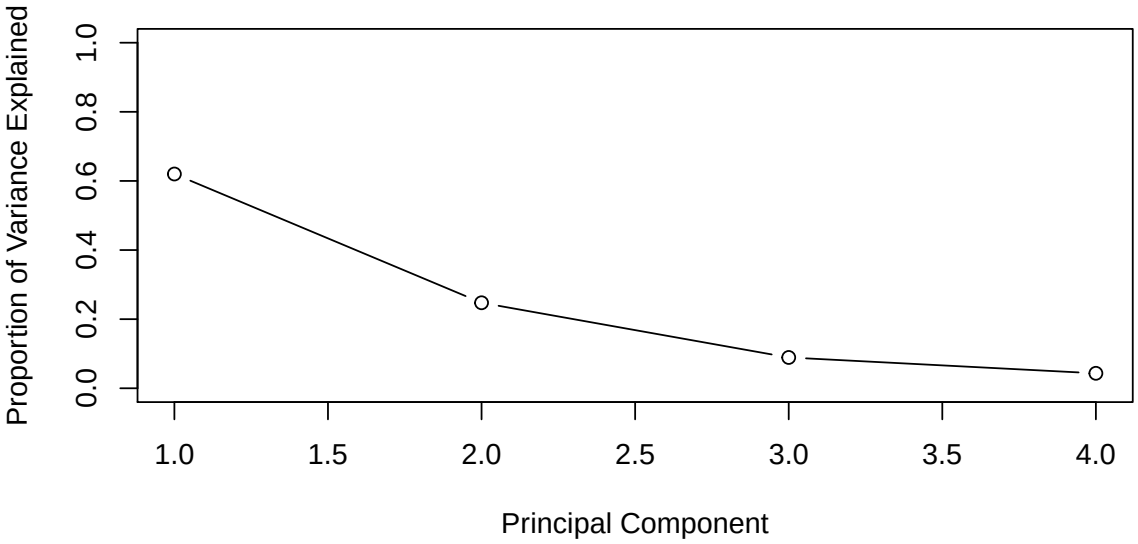


주성분의 사인을 바꾼것이다.

```
pr.out$sdev

## [1] 1.5748783 0.9948694 0.5971291 0.4164494

pr.var=pr.out$sdev^2
pr.var

## [1] 2.4802416 0.9897652 0.3565632 0.1734301

pve=pr.var/sum(pr.var)
pve

## [1] 0.62006039 0.24744129 0.08914080 0.04335752
```

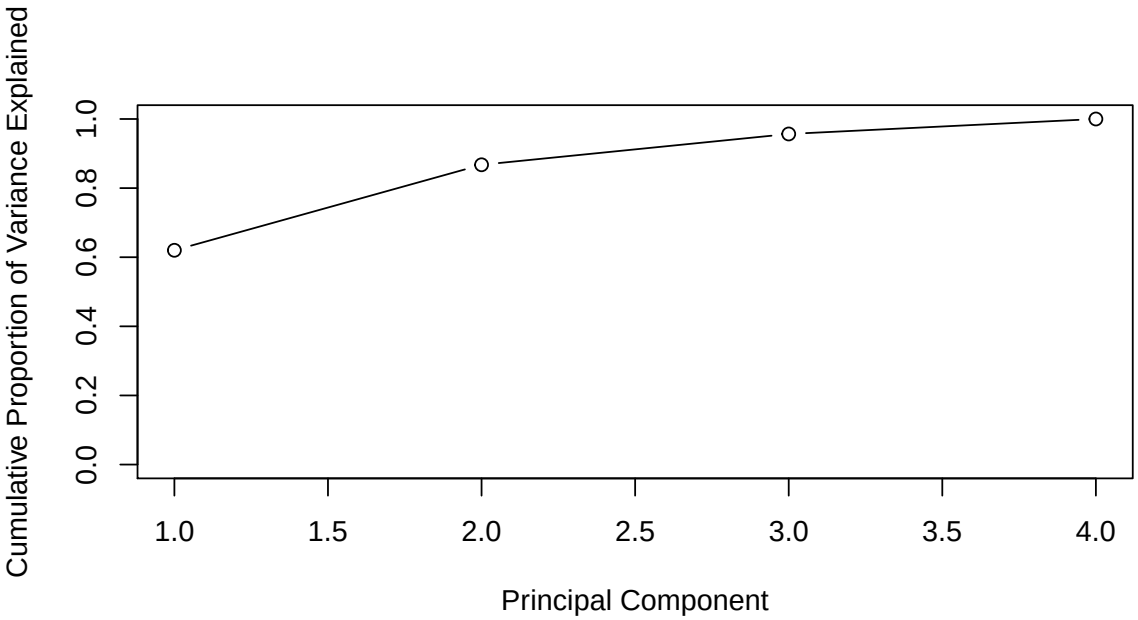
sdev는 주성분의 표준편차이다. 주성분의 분산을 계산하였다. 각 주성분의 설명 비율이다.

```
plot(pve, xlab="Principal Component", ylab="Proportion of Variance Explained", ylim=c(0,1),type='b')
```



각 주성분의 설명하는 분산 비율의 그림이다.

```
plot(cumsum(pve), xlab="Principal Component", ylab="Cumulative Proportion of Variance Explained", ylim=
```



주성분의 누적 분산 비율이다.

3 주성분회귀분석

```
library(pls)
```

```
##
## Attaching package: 'pls'
##
## The following object is masked from 'package:stats':
##
##    loadings
```

Partial Least Squares Regression (PLSR), Principal Component Regression (PCR) and Canonical Powered Partial Least Squares (CPPLS)가 있는 패키지이다.

```
set.seed(2)
str(Hitters)
```

```
## 'data.frame': 322 obs. of  20 variables:
##  $ AtBat    : int  293 315 479 496 321 594 185 298 323 401 ...
##  $ Hits     : int  66 81 130 141 87 169 37 73 81 92 ...
##  $ HmRun    : int  1 7 18 20 10 4 1 0 6 17 ...
##  $ Runs     : int  30 24 66 65 39 74 23 24 26 49 ...
##  $ RBI      : int  29 38 72 78 42 51 8 24 32 66 ...
##  $ Walks    : int  14 39 76 37 30 35 21 7 8 65 ...
##  $ Years    : int  1 14 3 11 2 11 2 3 2 13 ...
##  $ CAtBat   : int  293 3449 1624 5628 396 4408 214 509 341 5206 ...
##  $ CHits    : int  66 835 457 1575 101 1133 42 108 86 1332 ...
##  $ CHmRun   : int  1 69 63 225 12 19 1 0 6 253 ...
##  $ CRuns    : int  30 321 224 828 48 501 30 41 32 784 ...
##  $ CRBI     : int  29 414 266 838 46 336 9 37 34 890 ...
##  $ CWalks   : int  14 375 263 354 33 194 24 12 8 866 ...
##  $ League   : Factor w/ 2 levels "A","N": 1 2 1 2 2 1 2 1 2 1 ...
##  $ Division : Factor w/ 2 levels "E","W": 1 2 2 1 1 2 1 2 2 1 ...
##  $ PutOuts  : int  446 632 880 200 805 282 76 121 143 0 ...
##  $ Assists  : int  33 43 82 11 40 421 127 283 290 0 ...
##  $ Errors   : int  20 10 14 3 4 25 7 9 19 0 ...
```

```
##  $ Salary   : num  NA 475 480 500 91.5 750 70 100 75 1100 ...
##  $ NewLeague: Factor w/ 2 levels "A","N": 1 2 1 2 2 1 1 1 2 1 ...
```

```
head(Hitters)
```

```
##                   AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits
## -Andy Allanson      293   66     1   30  29    14     1    293    66
## -Alan Ashby         315   81     7   24  38    39    14   3449   835
## -Alvin Davis        479  130    18   66  72    76     3   1624   457
## -Andre Dawson       496  141    20   65  78    37    11   5628  1575
## -Andres Galarraga   321   87    10   39  42    30     2    396   101
## -Alfredo Griffin    594  169     4   74  51    35    11   4408  1133
##                   CHmRun CRuns CRBI CWalks League Division PutOuts Assists
## -Andy Allanson         1    30   29     14      A        E     446      33
## -Alan Ashby           69   321  414    375      N        W     632      43
## -Alvin Davis          63   224  266    263      A        W     880      82
## -Andre Dawson        225   828  838    354      N        E     200      11
## -Andres Galarraga     12    48   46     33      N        E     805      40
## -Alfredo Griffin      19   501  336    194      A        W     282     421
##                   Errors Salary NewLeague
## -Andy Allanson        20     NA         A
## -Alan Ashby           10  475.0         N
## -Alvin Davis          14  480.0         A
## -Andre Dawson          3  500.0         N
## -Andres Galarraga      4   91.5         N
## -Alfredo Griffin      25  750.0         A
```

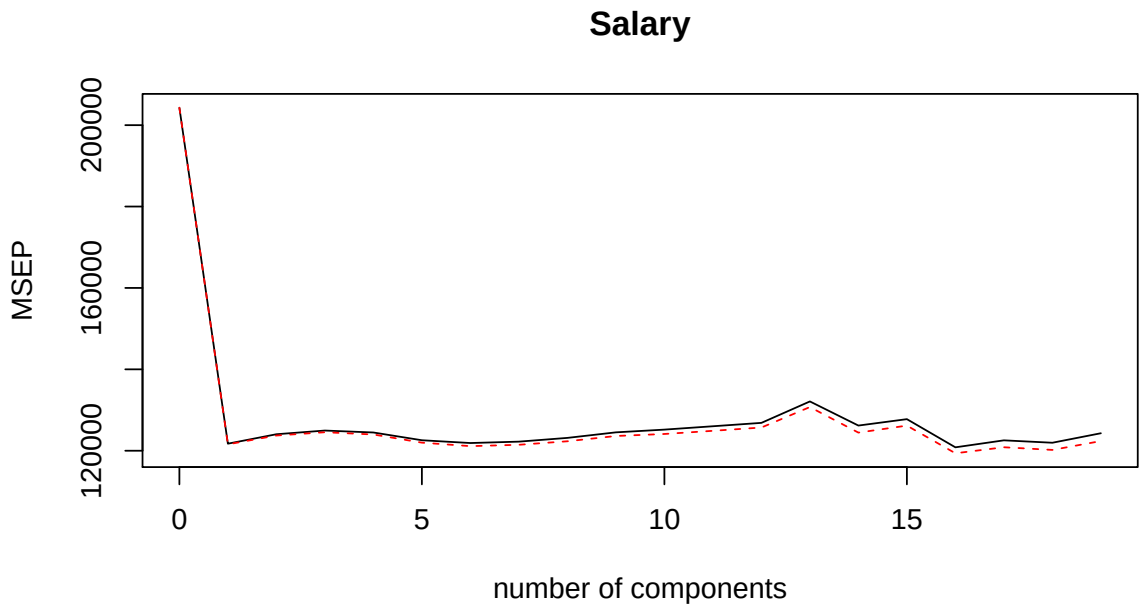
```
pcr.fit=pcr(Salary~., data=Hitters,scale=TRUE,validation="CV")
summary(pcr.fit)
```

```
## Data:  X dimension: 263 19
##  Y dimension: 263 1
## Fit method: svdpc
## Number of components considered: 19
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
```

```
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV             452    348.9    352.2    353.5    352.8    350.1    349.1
## adjCV          452    348.7    351.8    352.9    352.1    349.3    348.0
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV       349.6    350.9    352.9     353.8     355.0     356.2     363.5
## adjCV    348.5    349.8    351.6     352.3     353.4     354.5     361.6
##        14 comps  15 comps  16 comps  17 comps  18 comps  19 comps
## CV        355.2     357.4     347.6     350.1     349.2     352.6
## adjCV     352.8     355.2     345.5     347.6     346.7     349.8
##
## TRAINING: % variance explained
##         1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X         38.31    60.16    70.84    79.03    84.29    88.63    92.26
## Salary    40.63    41.58    42.17    43.22    44.90    46.48    46.69
##         8 comps  9 comps  10 comps  11 comps  12 comps  13 comps  14 comps
## X         94.96    96.28     97.26     97.98     98.65     99.15     99.47
## Salary    46.75    46.86     47.76     47.82     47.85     48.10     50.40
##         15 comps  16 comps  17 comps  18 comps  19 comps
## X          99.75     99.89     99.97     99.99    100.00
## Salary     50.55     53.01     53.85     54.61     54.61
```

Hitters는 타자들의 연봉에 대한 자료로 322명의 타자들에 대한 20개의 변수를 포함하였다. 함수 pcr의 사용법은 lm 함수와 비슷하다. scale은 설명변수를 척도화해서 주성분분석을 돌리는 것이다. validation="CV"는 주성분의 개수를 정하는데 10겹 교차검증을 사용하게 한다.

```
validationplot(pcr.fit,val.type="MSEP")
```



교차검증에러를 그림으로 그린 것이다.

```
set.seed(1)
pcr.fit=pcr(Salary~., data=Hitters,subset=train,scale=TRUE, validation="CV")

validationplot(pcr.fit,val.type="MSEP")
```

```
pcr.pred=predict(pcr.fit,x[test,],ncomp=7)

mean((pcr.pred-y.test)^2)

pcr.fit=pcr(y~x,scale=TRUE,ncomp=7)

summary(pcr.fit)

## Data:  X dimension: 263 19
##  Y dimension: 263 1
## Fit method: svdpc
## Number of components considered: 19
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV             452    348.9    352.2    353.5    352.8    350.1    349.1
## adjCV          452    348.7    351.8    352.9    352.1    349.3    348.0
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV       349.6    350.9    352.9     353.8     355.0     356.2     363.5
## adjCV    348.5    349.8    351.6     352.3     353.4     354.5     361.6
##        14 comps  15 comps  16 comps  17 comps  18 comps  19 comps
## CV        355.2     357.4     347.6     350.1     349.2     352.6
## adjCV     352.8     355.2     345.5     347.6     346.7     349.8
##
## TRAINING: % variance explained
##         1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X         38.31    60.16    70.84    79.03    84.29    88.63    92.26
## Salary    40.63    41.58    42.17    43.22    44.90    46.48    46.69
##         8 comps  9 comps  10 comps  11 comps  12 comps  13 comps  14 comps
## X         94.96    96.28     97.26     97.98     98.65     99.15     99.47
## Salary    46.75    46.86     47.76     47.82     47.85     48.10     50.40
##         15 comps  16 comps  17 comps  18 comps  19 comps
## X          99.75     99.89     99.97     99.99    100.00
## Salary     50.55     53.01     53.85     54.61     54.61
```

훈련자료에 주성분회귀분석을 적합하고 시험오차를 계산하였다.