Class Activity 12

Courtney Hodge

2024-08-05

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.3     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.4     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
library(pls)
```

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

```
baseball <- read.csv( "C:\\Users\\hodge\\Downloads\\Baseball.csv")
```

```
baseball <- na.omit(baseball)
```

1a

> in comparison to OLS, PCR lowers the risk of multicolinearity within our model and it shrinks the the # of columns to containt the same information as the *p* numerical random variable, but explained in a more convenient way.

1b

```
colnames(baseball)
```

```
##  [1] "X"         "AtBat"     "Hits"      "HmRun"     "Runs"      "RBI"
##  [7] "Walks"     "Years"     "CAtBat"    "CHits"     "CHmRun"    "CRuns"
## [13] "CRBI"      "CWalks"    "League"    "Division"  "PutOuts"   "Assists"
## [19] "Errors"    "Salary"    "NewLeague"
```

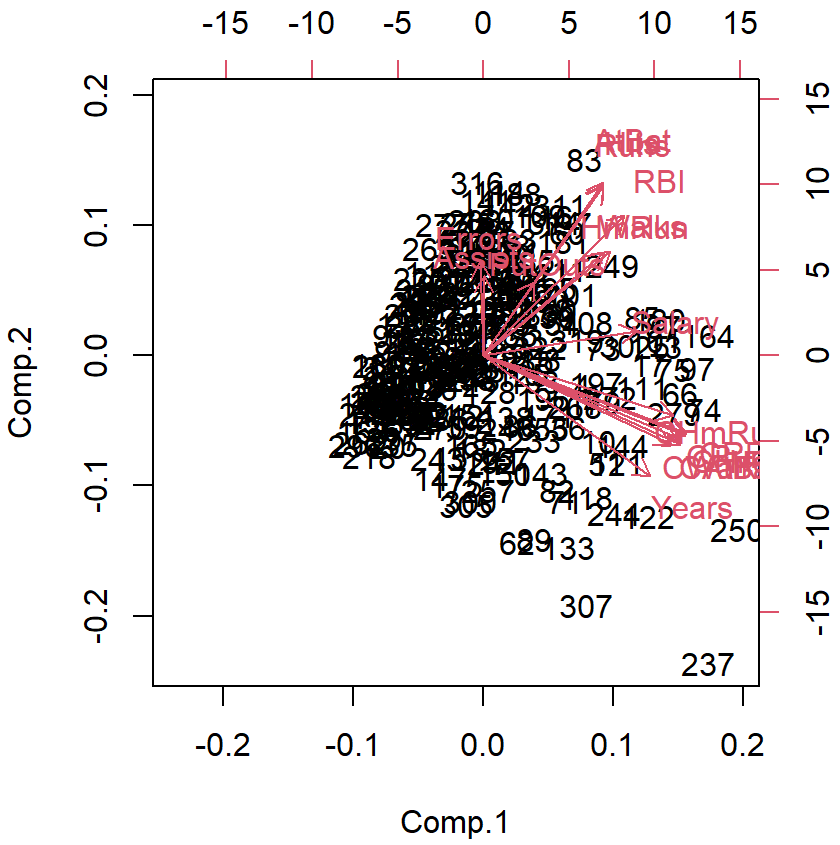
```
baseball_subset <- subset(baseball, select = -c(X, League, Division, NewLeague))

pca <- princomp(baseball_subset, fix_sign = T, cor = T)

summary(pca)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
## Standard deviation     2.7733967 2.0302601 1.3148557 0.9575410 0.84109683
## Proportion of Variance 0.4524547 0.2424680 0.1016968 0.0539344 0.04161435
## Cumulative Proportion  0.4524547 0.6949227 0.7966195 0.8505539 0.89216822
##                            Comp.6     Comp.7     Comp.8     Comp.9     Comp.10
## Standard deviation     0.72374220 0.69841796 0.50090065 0.42525940 0.363901982
## Proportion of Variance 0.03081193 0.02869339 0.01475891 0.01063797 0.007789685
## Cumulative Proportion  0.92298014 0.95167354 0.96643244 0.97707042 0.984860104
##                           Comp.11     Comp.12     Comp.13     Comp.14
## Standard deviation     0.312011679 0.243641510 0.232044829 0.163510472
## Proportion of Variance 0.005726546 0.003491834 0.003167341 0.001572687
## Cumulative Proportion  0.990586651 0.994078485 0.997245826 0.998818513
##                            Comp.15      Comp.16      Comp.17
## Standard deviation     0.1186398422 0.0693395039 3.466841e-02
## Proportion of Variance 0.0008279654 0.0002828216 7.069994e-05
## Cumulative Proportion  0.9996464785 0.9999293001 1.000000e+00
```

```
biplot(pca)
```



1c

PCRRRRR

```
pcareg <-pcr(Salary~., data = baseball_subset, scale = T, ) #does principal component regression
```

1d

> based on the model summary below, the R^2 if we used the first 3 principal components looks to be 41.93, and the R^2 if we used the first 10 principal components is 45.67.

```
summary(pcareg)
```

```
## Data:    X dimension: 263 16
##  Y dimension: 263 1
## Fit method: svdpc
## Number of components considered: 16
## TRAINING: % variance explained
##         1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
## X         45.31    71.00    81.80    87.24    91.60    94.80    96.37    97.53
## Salary    40.67    41.87    41.93    43.96    44.36    44.43    44.53    45.36
##         9 comps  10 comps  11 comps  12 comps  13 comps  14 comps  15 comps
## X         98.36     98.97     99.35     99.70     99.87     99.96     99.99
## Salary    45.36     45.67     47.59     48.36     50.81     51.97     52.78
##         16 comps
## X         100.00
## Salary     52.79
```

1e

> The salary prediction for the first two rows of the dataset using the first 3 principal components is 509.4619 for the 2nd row and 634.2457 for the third row.

```
new_dat = baseball_subset[1:2, ]
predict(pcareg, new_dat, ncomp = 3)
```

```
## , , 3 comps
##
##     Salary
## 2 509.4619
## 3 634.2457
```

> and with the first 10 principal components, we have salary predictions of 507.9765 for the 2nd row and 763.0485 for the 3rd row.

```
new_dat = baseball_subset[1:2, ]
predict(pcareg, new_dat, ncomp = 10)
```

```
## , , 10 comps
##
##     Salary
## 2 507.9765
## 3 763.0485
```

2

> some advantages for using a Lasso regression instead of PCR are the feature selection that automatically selects the most important features in a dataset. PCR considers all input predictors.