lab 06 PCR Solution

Principal Components Regression (PCR)

Load data & remove NA

```
library(ISLR)
library(pls)
```

```
##
## Attaching package: 'pls'
```

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

```
str(Hitters, vec.len = 1)
```

```
## 'data.frame':    322 obs. of  20 variables:
##  $ AtBat    : int  293 315 ...
##  $ Hits     : int  66 81 ...
##  $ HmRun    : int  1 7 ...
##  $ Runs     : int  30 24 ...
##  $ RBI      : int  29 38 ...
##  $ Walks    : int  14 39 ...
##  $ Years    : int  1 14 ...
##  $ CAtBat   : int  293 3449 ...
##  $ CHits    : int  66 835 ...
##  $ CHmRun   : int  1 69 ...
##  $ CRuns    : int  30 321 ...
##  $ CRBI     : int  29 414 ...
##  $ CWalks   : int  14 375 ...
##  $ League   : Factor w/ 2 levels "A","N": 1 2 ...
##  $ Division : Factor w/ 2 levels "E","W": 1 2 ...
##  $ PutOuts  : int  446 632 ...
##  $ Assists  : int  33 43 ...
##  $ Errors   : int  20 10 ...
##  $ Salary   : num  NA 475 ...
##  $ NewLeague: Factor w/ 2 levels "A","N": 1 2 ...
```

```
Hitters <- na.omit(Hitters)
Hitters_mat <- model.matrix(Salary ~ ., Hitters)
X <- scale(Hitters_mat[, -1], center = T, scale = T)
Y <- Hitters$Salary
```

2.1) Start with PCA

You can also embed plots, for example:

```
# principal component regression (without CV)
pcr_fit <- pcr(Salary ~ ., data = Hitters, scale = TRUE, validation = "none")
names(pcr_fit)
```

```
##  [1] "coefficients"  "scores"        "loadings"      "Yloadings"
##  [5] "projection"    "Xmeans"        "Ymeans"        "fitted.values"
##  [9] "residuals"     "Xvar"          "Xtotvar"       "fit.time"
## [13] "ncomp"         "method"        "scale"         "call"
## [17] "terms"         "model"
```

```
scores <- matrix(scores(pcr_fit), nrow = NROW(X), ncol = NCOL(X))
svd_Hitters <- svd(X)
svd_Z <- X %*% svd_Hitters$v
```

2.2) PC Regression on the first component

The fitted response using PC1 provided by pcr() is the same as the result using svd() and some linear algebra:

```
z1 <- svd_Z[, 1]
b1 <- as.numeric(solve(t(z1) %*% z1) %*% t(z1) %*% Y)
yhat_PC1 <- as.vector(b1 * z1) + mean(Y)
fitted_pcr_PC1 <- as.vector(pcr_fit$fitted.values[ , , 1])
all.equal(yhat_PC1, fitted_pcr_PC1)
```

```
## [1] TRUE
```

2.3) PC Regression on all PCs

```
bPCR <- solve(t(svd_Z) %*% svd_Z) %*% t(svd_Z) %*% Y
yhat_full <- as.numeric(svd_Z %*% bPCR + mean(Y))
fitted_pcr_full <- as.numeric(pcr_fit$fitted.values[ , , 19])
all.equal(yhat_full, fitted_pcr_full)
```

```
## [1] TRUE
```

2.4) PCR coefficients in terms of the predictor variables

i.e., TO verify if $V^T \beta_{OLS} = \beta_{PCR}$

```
# First PC
b1_star <- b1 * svd_Hitters$v[, 1]
coef_pcr_PC1 <- as.numeric(pcr_fit$coefficients[ , , 1])
all.equal(b1_star, coef_pcr_PC1)
```

```
## [1] TRUE
```

```
# First and second PCs
b12_star <- as.numeric(svd_Hitters$v[, 1:2] %*% solve(diag((svd_Hitters$d[1:2])))
                       %*% t(svd_Hitters$u[, 1:2]) %*% Y)
coef_pcr_PC12 <- as.numeric(pcr_fit$coefficients[ , , 2])
all.equal(b12_star, coef_pcr_PC12)
```

```
## [1] TRUE
```

```
# Alternative (sum them up)
z2 <- svd_Z[, 2]
b2 <- as.numeric(solve(t(z2) %*% z2) %*% t(z2) %*% Y)
```

```r
b2_star <- b2 * svd_Hitters$v[, 2]
all.equal(b12_star, b1_star + b2_star)
```

```
## [1] TRUE
```

All possible sets of PCs

```r
for (i in 2:19) {
  b_star <- as.numeric(svd_Hitters$v[, 1:i] %*% solve(diag((svd_Hitters$d[1:i])))
                       %*% t(svd_Hitters$u[, 1:i]) %*% Y)
  coef_pcr <- as.numeric(pcr_fit$coefficients[ , , i])
  print(paste0("when using ", i, " PCs, the comparison is ",
               all.equal(b_star, coef_pcr)))
}
```

```
## [1] "when using 2 PCs, the comparison is TRUE"
## [1] "when using 3 PCs, the comparison is TRUE"
## [1] "when using 4 PCs, the comparison is TRUE"
## [1] "when using 5 PCs, the comparison is TRUE"
## [1] "when using 6 PCs, the comparison is TRUE"
## [1] "when using 7 PCs, the comparison is TRUE"
## [1] "when using 8 PCs, the comparison is TRUE"
## [1] "when using 9 PCs, the comparison is TRUE"
## [1] "when using 10 PCs, the comparison is TRUE"
## [1] "when using 11 PCs, the comparison is TRUE"
## [1] "when using 12 PCs, the comparison is TRUE"
## [1] "when using 13 PCs, the comparison is TRUE"
## [1] "when using 14 PCs, the comparison is TRUE"
## [1] "when using 15 PCs, the comparison is TRUE"
## [1] "when using 16 PCs, the comparison is TRUE"
## [1] "when using 17 PCs, the comparison is TRUE"
## [1] "when using 18 PCs, the comparison is TRUE"
## [1] "when using 19 PCs, the comparison is TRUE"
```

```r
# bOLS <- solve(t(X) %*% X) %*% t(X) %*% Y
# all.equal(t(svd_Hitters$v) %*% bOLS, bPCR)
```