27) Principal Components Regression, and Partial Least Squares

Vitor Kamada

February 2018

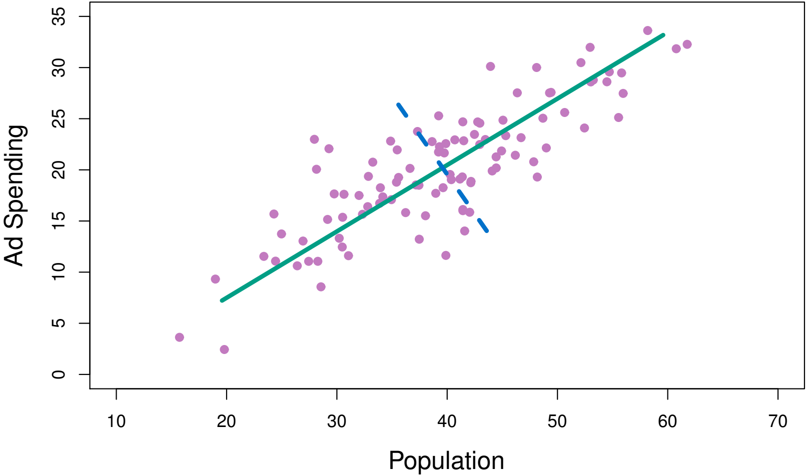
Reference

Tables, Graphics, and Figures from

An Introduction to Statistical Learning

James et al. (2017): Chapters: 6.3, and 6.7

Advertising Data



Dimension Reduction

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$$

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i$$

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{jm} x_{ij}$$

$$= \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{jm} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij}$$

Principal Components Analysis (PCA)

$$Z_1 = \phi_{11}(pop - \bar{pop}) + \phi_{21}(ad - \bar{ad})$$

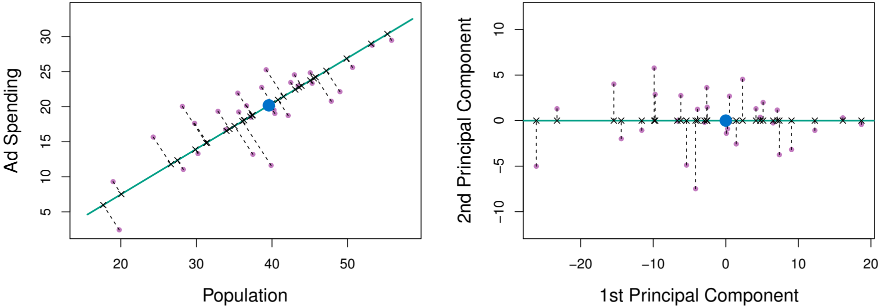
$$Z_1 = 0.839(pop - \bar{pop}) + 0.544(ad - \bar{ad})$$

$$Var[\phi_{11}(pop - \bar{pop}) + \phi_{21}(ad - \bar{ad})]$$

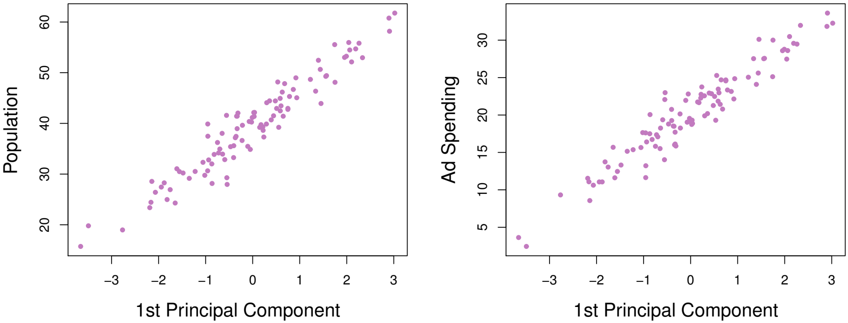
$$\phi_{11}^2 + \phi_{21}^2 = 1$$

$$z_{i1} = 0.839(pop_i - \bar{pop}) + 0.544(ad_i - \bar{ad})$$

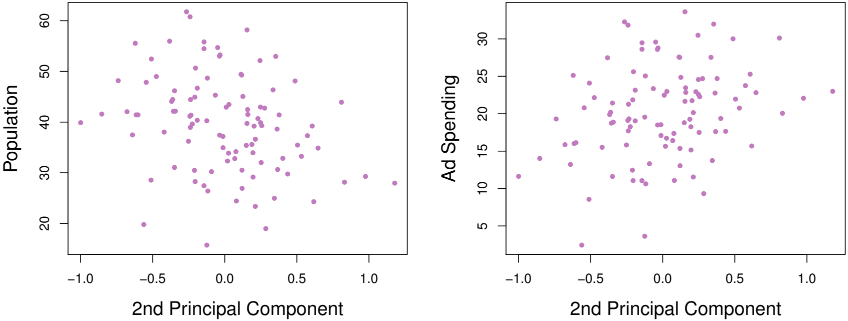
First and Second Principal Component



First Principal Component Scores z_{i1} vs pop and ad

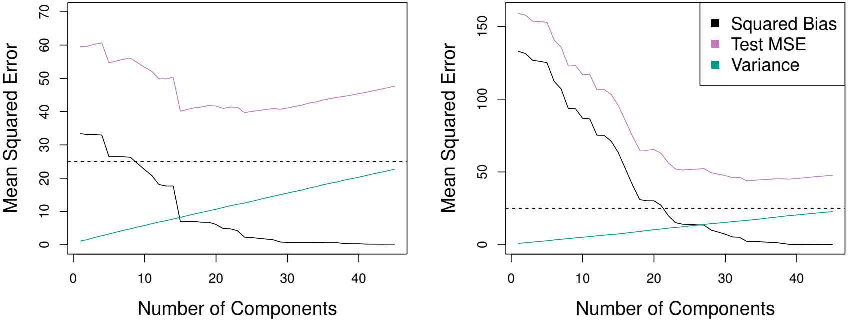


Second Principal Component Scores z_{i2} vs pop and ad



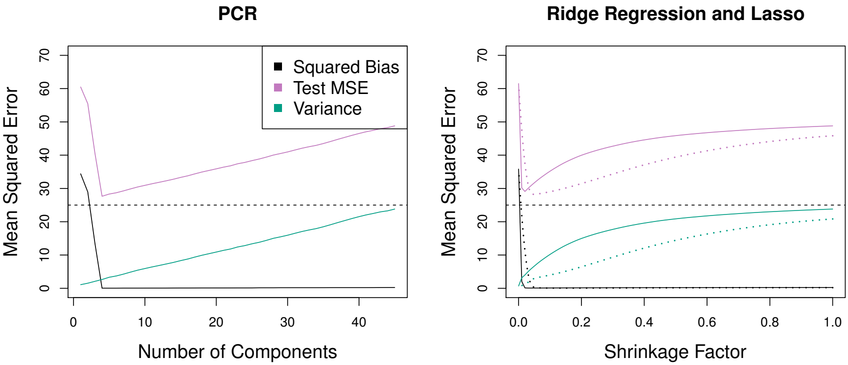
Principal Components Regression (PCR) for Simulated Data

Horizontal Dashed Line: $Var(\epsilon)$



Simulated Data in which the first 5 PC of X contain all the information about Y

Solid (lasso), Dotted(ridge)

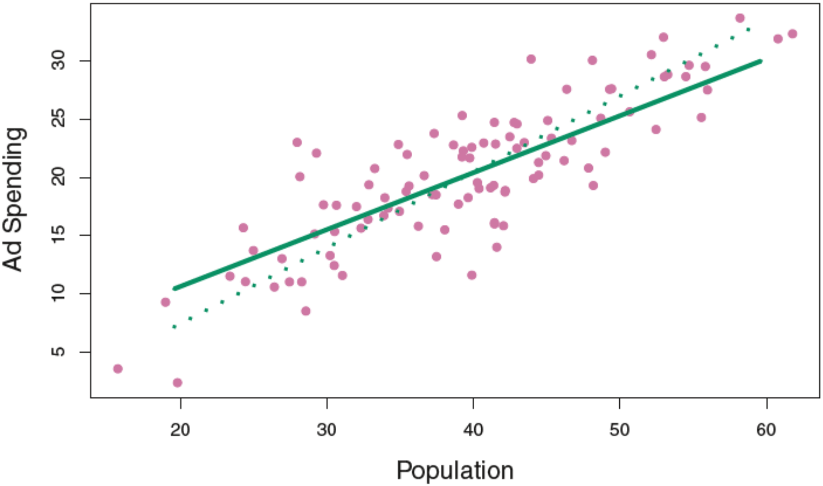


Partial Least Squares (PLS)

In computing $Z_1 = \sum_{j=1}^{p} \phi_{j1} X_j$, PLS places the highest weight on the variables that are most strongly related to Y

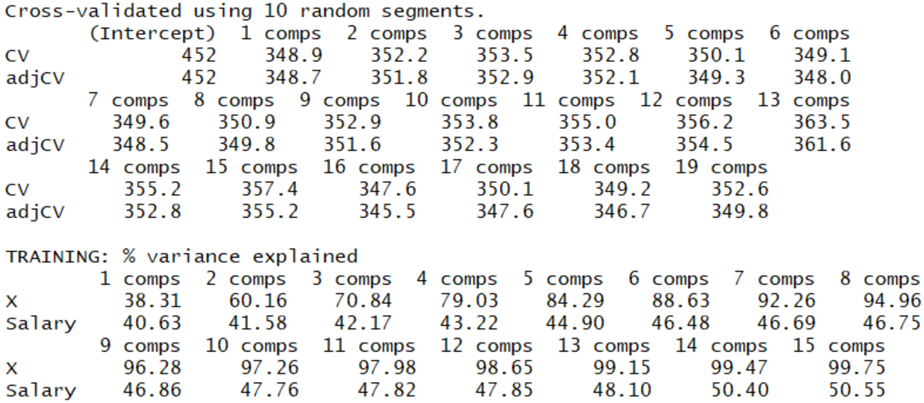
First PLS Direction (solid line) and First PCR Direction (dotted line)

$Y = \text{Sales}$



library (pls); set.seed (2)

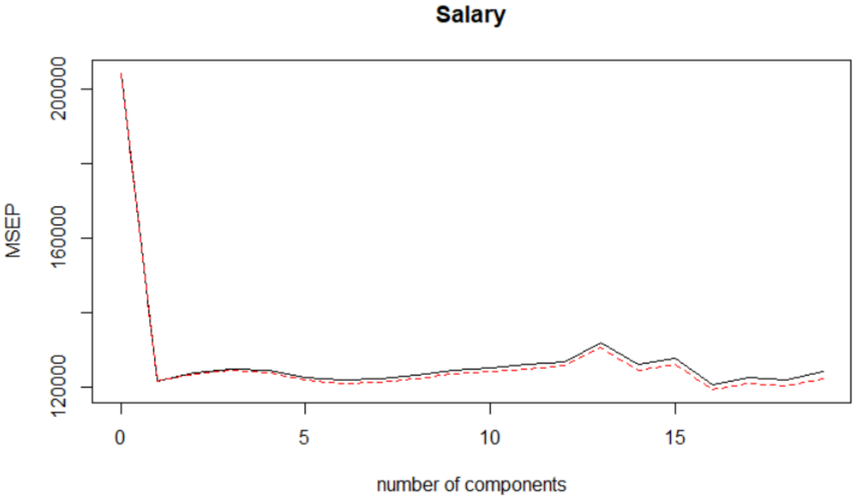
pcr.fit=pcr(Salary~., data=Hitters, scale=TRUE,validation="CV"); summary(pcr.fit)



```
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  8 comps
X         38.31    60.16    70.84    79.03    84.29    88.63    92.26    94.96
Salary    40.63    41.58    42.17    43.22    44.90    46.48    46.69    46.75
        9 comps  10 comps  11 comps  12 comps  13 comps  14 comps  15 comps
X         96.28     97.26     97.98     98.65     99.15     99.47     99.75
Salary    46.86     47.76     47.82     47.85     48.10     50.40     50.55
```

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

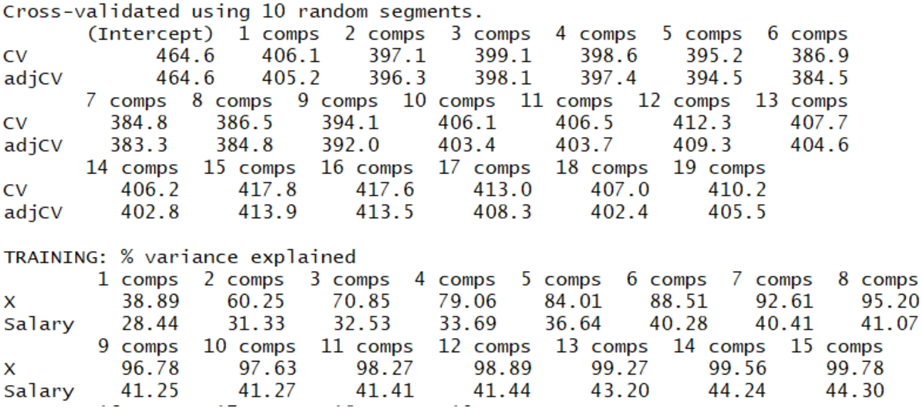


Training and Test Data Set

```
x=model.matrix(Salary~.,Hitters)[,-1]
y=Hitters$Salary; set.seed(1)
train=sample(1:nrow(x), nrow(x)/2)
test=(-train); y.test=y[test]

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

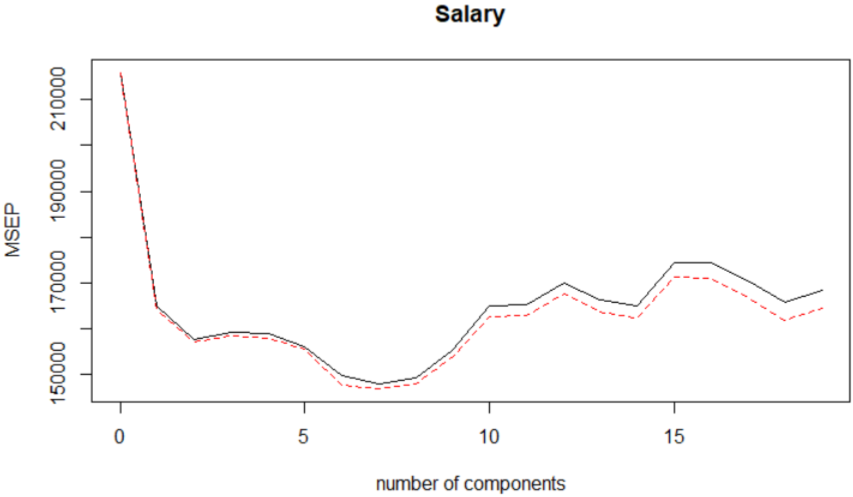
summary(pcr.fit)



```
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV           464.6    406.1    397.1    399.1    398.6    395.2    386.9
adjCV        464.6    405.2    396.3    398.1    397.4    394.5    384.5
       7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
CV       384.8    386.5    394.1     406.1     406.5     412.3     407.7
adjCV    383.3    384.8    392.0     403.4     403.7     409.3     404.6
       14 comps  15 comps  16 comps  17 comps  18 comps  19 comps
CV        406.2     417.8     417.6     413.0     407.0     410.2
adjCV     402.8     413.9     413.5     408.3     402.4     405.5

TRAINING: % variance explained
        1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
X         38.89    60.25    70.85    79.06    84.01    88.51    92.61    95.20
Salary    28.44    31.33    32.53    33.69    36.64    40.28    40.41    41.07
        9 comps  10 comps  11 comps  12 comps  13 comps  14 comps  15 comps
X         96.78     97.63     98.27     98.89     99.27     99.56     99.78
Salary    41.25     41.27     41.41     41.44     43.20     44.24     44.30
```

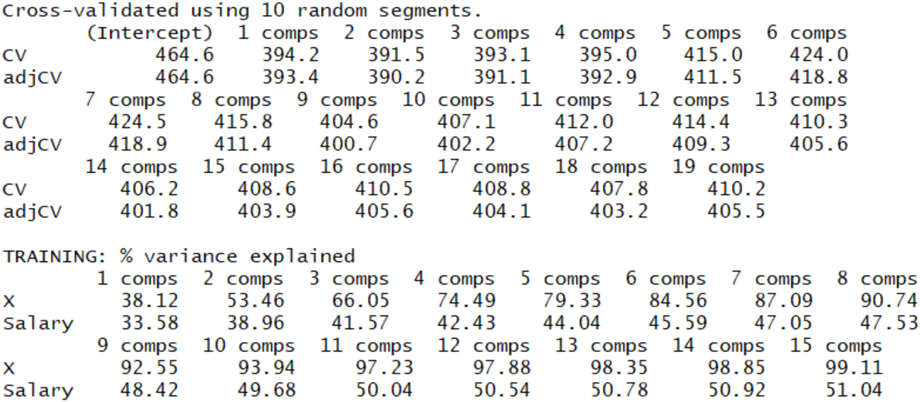
validationplot(pcr.fit,val.type="MSEP")



Partial Least Squares (PLS)

set.seed(1)

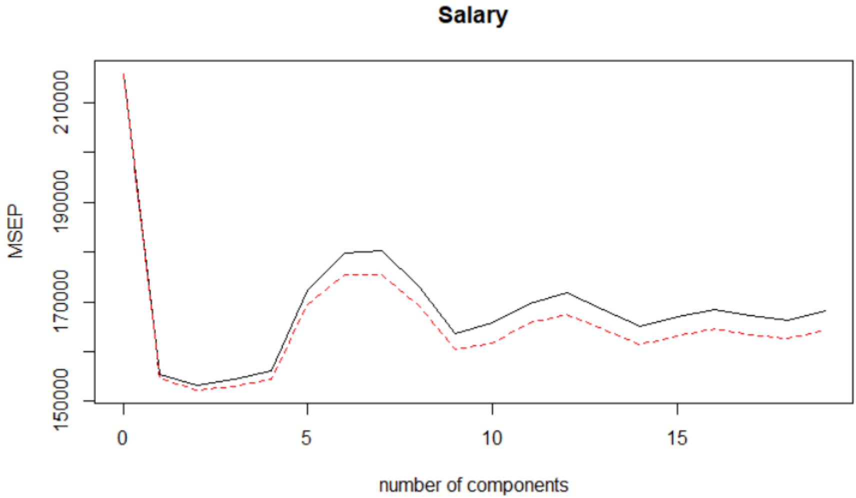
pls.fit=plsr(Salary∼., data=Hitters, subset=train, scale=TRUE, validation="CV"); summary(pls.fit)



```
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV           464.6    394.2    391.5    393.1    395.0    415.0    424.0
adjCV        464.6    393.4    390.2    391.1    392.9    411.5    418.8
       7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
CV       424.5    415.8    404.6     407.1     412.0     414.4     410.3
adjCV    418.9    411.4    400.7     402.2     407.2     409.3     405.6
       14 comps  15 comps  16 comps  17 comps  18 comps  19 comps
CV        406.2     408.6     410.5     408.8     407.8     410.2
adjCV     401.8     403.9     405.6     404.1     403.2     405.5

TRAINING: % variance explained
        1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
X         38.12    53.46    66.05    74.49    79.33    84.56    87.09    90.74
Salary    33.58    38.96    41.57    42.43    44.04    45.59    47.05    47.53
        9 comps  10 comps  11 comps  12 comps  13 comps  14 comps  15 comps
X         92.55     93.94     97.23     97.88     98.35     98.85     99.11
Salary    48.42     49.68     50.04     50.54     50.78     50.92     51.04
```

validationplot(pls.fit,val.type="MSEP")


Salary
MSEP
number of components