HW5

2019150445/Shin Baek Rok

2020 12 13

```r
#load data
raw.data<-read.table('https://learn-ap-northeast-2-prod-fleet01-xythos.content.blackboardcdn.com/5d3914
glimpse(raw.data)
```

```
## Rows: 30
## Columns: 12
## $ Y   <dbl> 18.9, 17.0, 20.0, 18.3, 20.1, 11.2, 22.1, 21.5, 34.7, 30.4, 16....
## $ X1  <dbl> 350.0, 350.0, 250.0, 351.0, 225.0, 440.0, 231.0, 262.0, 89.7, 9...
## $ X2  <int> 165, 170, 105, 143, 95, 215, 110, 110, 70, 75, 155, 80, 109, 11...
## $ X3  <int> 260, 275, 185, 255, 170, 330, 175, 200, 81, 83, 250, 83, 146, 1...
## $ X4  <dbl> 8.00, 8.50, 8.25, 8.00, 8.40, 8.20, 8.00, 8.50, 8.20, 9.00, 8.5...
## $ X5  <dbl> 2.56, 2.56, 2.73, 3.00, 2.76, 2.88, 2.56, 2.56, 3.90, 4.30, 3.0...
## $ X6  <int> 4, 4, 1, 2, 1, 4, 2, 2, 2, 2, 4, 2, 2, 1, 2, 2, 4, 4, 4, 2, 2, ...
## $ X7  <int> 3, 3, 3, 3, 3, 3, 3, 3, 4, 5, 3, 4, 4, 3, 4, 3, 3, 3, 3, 3, 3, ...
## $ X8  <dbl> 200.3, 199.6, 196.7, 199.9, 194.1, 184.5, 179.3, 179.3, 155.7, ...
## $ X9  <dbl> 69.9, 72.9, 72.2, 74.0, 71.8, 69.0, 65.4, 65.4, 64.0, 65.0, 74....
## $ X10 <int> 3910, 3860, 3510, 3890, 3365, 4215, 3020, 3180, 1905, 2320, 388...
## $ X11 <int> 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, ...
```

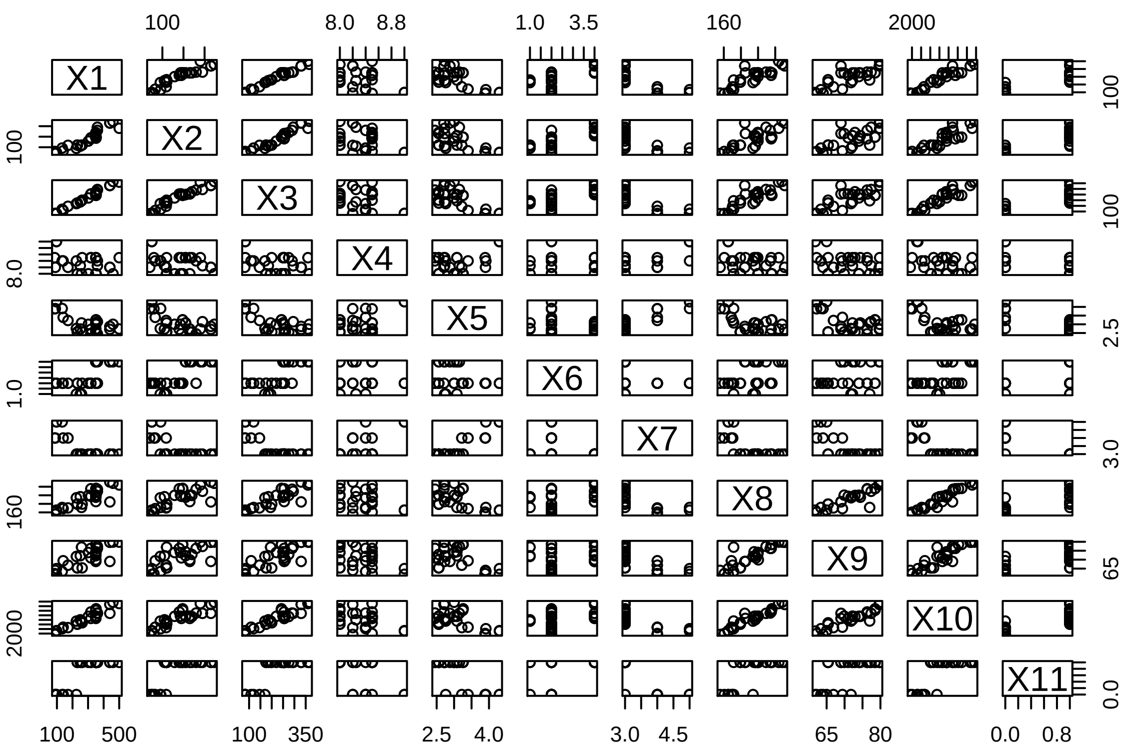
1.

```r
cor(raw.data[,-1])
```

```
##             X1         X2         X3          X4         X5          X6
## X1   1.0000000  0.9406456  0.9895851 -0.34958682 -0.6714311  0.63996417
## X2   0.9406456  1.0000000  0.9643592 -0.28989951 -0.5509642  0.76141897
## X3   0.9895851  0.9643592  1.0000000 -0.32599915 -0.6728661  0.65312630
## X4  -0.3495868 -0.2898995 -0.3259992  1.00000000  0.4137808  0.03748643
## X5  -0.6714311 -0.5509642 -0.6728661  0.41378081  1.0000000 -0.21952829
## X6   0.6399642  0.7614190  0.6531263  0.03748643 -0.2195283  1.00000000
## X7  -0.7717815 -0.6259445 -0.7461800  0.55823570  0.8717662 -0.27563863
## X8   0.8649023  0.8027387  0.8641224 -0.30415026 -0.5613315  0.42206800
## X9   0.8001582  0.7105117  0.7881284 -0.37817358 -0.4534470  0.30038618
## X10  0.9531271  0.8878810  0.9434871 -0.35845879 -0.5798617  0.52036693
## X11  0.8241409  0.7086735  0.8012765 -0.44054570 -0.7546650  0.39548928
##             X7         X8         X9        X10        X11
## X1  -0.7717815  0.8649023  0.8001582  0.9531271  0.8241409
## X2  -0.6259445  0.8027387  0.7105117  0.8878810  0.7086735
## X3  -0.7461800  0.8641224  0.7881284  0.9434871  0.8012765
## X4   0.5582357 -0.3041503 -0.3781736 -0.3584588 -0.4405457
```

```
## X5   0.8717662 -0.5613315 -0.4534470 -0.5798617 -0.7546650
## X6  -0.2756386  0.4220680  0.3003862  0.5203669  0.3954893
## X7   1.0000000 -0.6552065 -0.6551300 -0.7058126 -0.8506963
## X8  -0.6552065  1.0000000  0.8831512  0.9554541  0.6824919
## X9  -0.6551300  0.8831512  1.0000000  0.8994711  0.6326677
## X10 -0.7058126  0.9554541  0.8994711  1.0000000  0.7530353
## X11 -0.8506963  0.6824919  0.6326677  0.7530353  1.0000000
```

```
plot(raw.data[,-1])
```



When we take a look at pairwise plot, we can find clear evidence of multicollinearity in X1&X2,X1&X3,X2&X3,X1&X5,X1&X8
Also in corr matrix, the absolute value of elements are big enough.

2.

```
cor.mat<-cor(raw.data[,-1])
eigenvalue<-eigen(cor.mat)$values
eigenvalue
```

```
##  [1] 7.702574847 1.403077880 0.773435643 0.577055424 0.211498935 0.141941470
##  [7] 0.095142049 0.050092536 0.033266309 0.008417705 0.003497202
```

```
eigenvalue[1]/sum(eigenvalue)
```

```
## [1] 0.7002341
```

```r
sum(eigenvalue[1:2])/sum(eigenvalue)
```

```
## [1] 0.8277866
```

```r
sum(eigenvalue[1:3])/sum(eigenvalue)
```

```
## [1] 0.8980989
```

```r
#We need three principle components to retain 85% of the information
```

3.

```r
lm(Y~.,data=raw.data) %>% car::vif()>10
```

```
##    X1    X2    X3    X4    X5    X6    X7    X8    X9   X10   X11
##  TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE
```

When we take VIF=10 as criteria, we can say X1,X2,X3,X7,X8,X10 are affected by the presence of multi-collinearity.

4.

```r
eigenvector<-eigen(cor.mat)$vector
U1<-as.matrix(eigenvector[,1])
U2<-as.matrix(eigenvector[,2])
X<-as.matrix(scale(raw.data[,-1]))
Z1<-X %*% U1
Z2<-X %*% U2
cbind(Z1,Z2) %>% head(3)#PC scores
```

```
##            [,1]       [,2]
## [1,] -1.7710783  0.1463517
## [2,] -1.7118429 -0.8936548
## [3,]  0.1762204  1.7017057
```

5.

```r
null=lm(Y~X6,data=raw.data)#Include X6
full=lm(Y~X1+X2+X3+X4+X5+X6+X7+X8+X9+X10, data=raw.data)#Exclude X11
step(null,scope=list(lower=null,upper=full),direction='forward',k=log(nrow(raw.data)))
```

```
## Start:  AIC=108.3
## Y ~ X6
##
##        Df Sum of Sq    RSS    AIC
## + X1    1    625.46 258.66  74.833
## + X3    1    580.81 303.31  79.610
## + X10   1    574.70 309.41  80.208
```

```
## + X2    1    517.43 366.69  85.303
## + X9    1    481.67 402.44  88.094
## + X8    1    423.56 460.56  92.141
## + X7    1    410.95 473.17  92.951
## + X5    1    337.75 546.36  97.266
## + X4    1    220.62 663.49 103.093
## <none>              884.12 108.304
##
## Step:  AIC=74.83
## Y ~ X6 + X1
##
##        Df Sum of Sq    RSS    AIC
## <none>              258.65 74.833
## + X4    1    9.4520 249.20 77.117
## + X3    1    6.2635 252.39 77.498
## + X9    1    4.6391 254.02 77.691
## + X8    1    2.2074 256.45 77.977
## + X10   1    1.0702 257.58 78.109
## + X5    1    0.8886 257.77 78.131
## + X7    1    0.0365 258.62 78.230
## + X2    1    0.0013 258.65 78.234
```

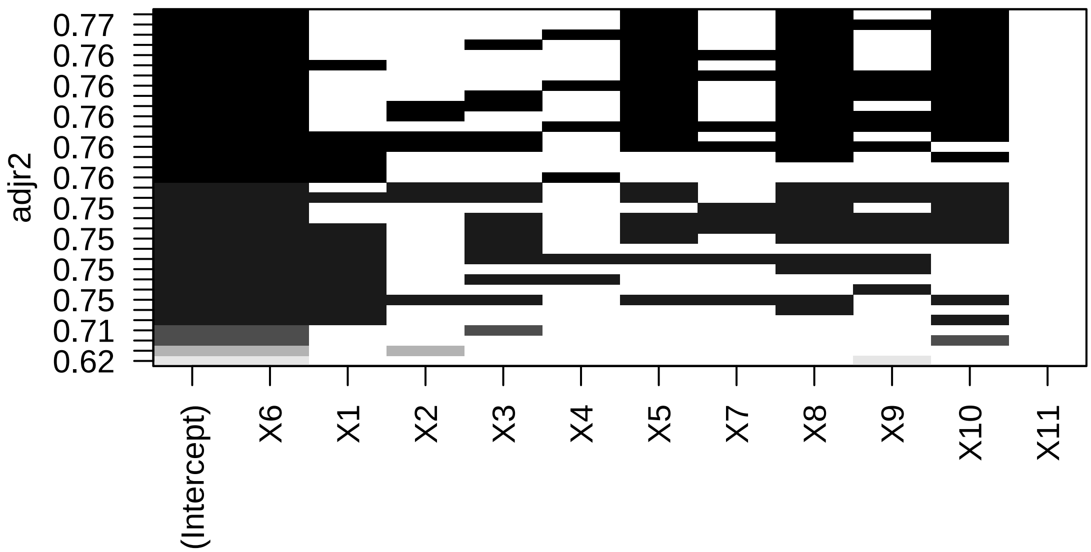
```
##
## Call:
## lm(formula = Y ~ X6 + X1, data = raw.data)
##
## Coefficients:
## (Intercept)           X6           X1
##    32.74244      0.85052     -0.05209
```

The fitted regression equation is

$$\hat{Y} = 32.74244 \quad - \quad 0.05209 \times X1 \quad + \quad 0.85052 \times X6$$

6.

```
Xi<-raw.data[,-1]
Y<-raw.data[,1]
leaps::regsubsets(Y~.,data=data.frame(Y,Xi),nbest=5,force.in=6,force.out=11) %>% plot(scale='adjr2')
```



X6,X5,X8,X10 are contained in my best model.

#7.

$M = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \overset{\text{let}}{=} Cov(X)$

To find eigenvalues, we set $Mx = \delta_1 I x$

$\Rightarrow (M - \delta_1 I)x = 0$

$\Rightarrow \begin{vmatrix} 2-\delta_1 & 1 \\ 1 & 2-\delta_1 \end{vmatrix} = 0$ Since X is not zero.

$\Leftrightarrow (2-\delta_1)(2-\delta_1) - 1 = 0$

$\Rightarrow \underbrace{\delta_1 = 3,\ \delta_2 = 1}_{\text{eigenvalues}}$ (set $\delta_1 \geq \delta_2$)

Consider $\delta_1 = 3$.

Let $u_1 = \begin{pmatrix} u_{11} \\ u_{21} \end{pmatrix}$ eigenvector where eigenvalue $= 3$.

$\Rightarrow (M - \delta_1 I)u_1 = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} u_{11} \\ u_{21} \end{pmatrix} = -u_{11} + u_{21} = 0$

$\Rightarrow u_{11} = u_{21}$

& we have $u_{11}^2 + u_{21}^2 = 1$ Since u_1 is unit vector.

$\therefore u_{11}^2 = \frac{1}{2},\ u_{21}^2 = \frac{1}{2}$

$\Rightarrow u_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$ or $\begin{pmatrix} -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix}$

Consider $\delta_2 = 1$

Let $u_2 = \begin{pmatrix} u_{12} \\ u_{22} \end{pmatrix}$ eigenvector where eigenvalue $= 1$

$\Rightarrow (M - \delta_2 I)u_2 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} u_{12} \\ u_{22} \end{pmatrix} = u_{12} + u_{22} = 0$

$\Rightarrow u_{12} = -u_{22}$

& $u_{12}^2 + u_{22}^2 = 1$

$\therefore u_2 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$ or $\begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix}$

Then, $z_1 = Xu_1 = (x_1\ x_2)\begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} = \frac{1}{\sqrt{2}}x_1 + \frac{1}{\sqrt{2}}x_2$

$z_2 = Xu_2 = (x_1\ x_2)\begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix} = \frac{1}{\sqrt{2}}x_1 - \frac{1}{\sqrt{2}}x_2$