多變量分析 作業一

109354003 統碩一 吳書恆

3/29/2021

Note: Please email your HW1 report (with name and ID, in pdf format) to the TA (108354007@nccu.edu.tw) before the deadline (13:00, 4/1/2021).

Topic: Principal Component Regression (PCR) and Partial Least Squares (PLS) regression

Data： “2020-QS-World-University-Rankings-100.csv” from the WM5 website. The data contains variables based on which the 2020 world college ranking was given by the QS. The variables are:

Rank: college ranking based on the “Overall_Score”
College Name: name of colleges from all places
Academic_Reputation (Y): score given for the college’ s overall academic reputation
Employer_Reputation (X1): score given by employers based on the quality of graduates
Faculty/Student (X2): score based on the college’ s faculty/student ratio
Faculty_Citation (X3): score based on citations per faculty
International_Faculty (X4): score based on the college’ s international faculty ratio
International_Students (X5): score based on the college’ s international student ratio
Overall_Score: score calculated based on the above category scores with weights

Question: Fit the following 3 regression models by using R and evaluate their performance in terms of the accuracy of predicting the response y = “Academic_Reputation” (using a 10-fold Cross Validation):

Model 1: The Least Squares (LS) regression model without the intercept term.
Model 2: The Principal Component Regression (PCR) without the intercept term. For this method, please choose the best number of components based on the model predictability.
Model 3: The Partial Least Squares (PLS) regression without the intercept term. Analogously, please choose the best number of components based on the model predictability.

(1) Are the above 3 prediction models similar, or different?
(2) Which model is best for predicting the college’ s “Academic Reputation” ? Explain why.

Result：(1) In order to compare three model, we need to decide the components to maintain first. In PCR and PLS, the Cross-validatation (CV) is listed, we find that 4 components and 2 components are prefered respectively since the lowest Root mean squared error (RMSE) in respective outputs. If the perpose is the explaination in response, we need to consider R-squred first. However, we usually consider prediction in PCR or PLS. **Table 2** only retain the 4 components and 3 components respectively in PCR and PLS and include 5 variables in LS.

Table 1 Cross-validated (RMSE) using 10 random segments.

	1 Comps	2 Comps	3 Comps	4 Comps	5 Comps
PCR	16.50	16.03	13.18	12.91	13.04
PLS	13.31	12.97	12.98	13.03	13.04

Table 2 Some comparisons (X variance explained, R-square, CV) in three models

	LS	PCR (4 Comps)	PLS (2 Comps)
X variance explained(%)	100	94.29	54.33
R-squred	0.4167	0.4162	0.4155
CV：RMSE	13.02	12.91	12.97

(2) PCR is best for predicting since the lowest RMSE in **Table 1**. However, for the dim-reduction performance, the PLS is better than PCR, which only 1-dim reducted. Although LS is not good at prediction, the imfomation of X can be maintained in model to explain.

Code:

```
> #data imput
> qs <- read.csv("2020-QS-World-University-Rankings-100.csv", header = T)
> #str(qs) #n=100, p=9
>
> #define model
> xnam <- names(qs)[4:8]
> (fmla <-  as.formula(paste("Academic_Reputation ~ ",
+                         paste(xnam, collapse= "+"))))
Academic_Reputation ~ Employer_Reputation + Faculty_Student + 
   Faculty_Citation + International_Faculty + International_Students
>
> #Least Squares (LS) regression model
> lm.fit <- lm(fmla, qs, y = TRUE, x = TRUE)
> summary(lm.fit)

Call:
lm(formula = fmla, data = qs, x = TRUE, y = TRUE)

Residuals:
    Min      1Q  Median      3Q     Max 
-48.066  -9.123   3.531   7.546  25.954 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)            38.122679   7.902173   4.824 5.41e-06 ***
Employer_Reputation     0.538521   0.067751   7.949 4.09e-12 ***
Faculty_Student         0.006841   0.048864   0.140    0.889    
Faculty_Citation        0.078606   0.059599   1.319    0.190    
International_Faculty  -0.055597   0.060695  -0.916    0.362    
International_Students -0.003385   0.062995  -0.054    0.957    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.8 on 94 degrees of freedom
Multiple R-squared:  0.4167,Adjusted R-squared:  0.3857 
F-statistic: 13.43 on 5 and 94 DF,  p-value: 7.026e-10

> library(lmvar)
> set.seed(2)
> cv.lm(lm.fit)
Mean absolute error        :  10.75483 
Sample standard deviation  :  2.1697 

Mean squared error         :  178.2395 
```

```
Sample standard deviation  :  87.04529

Root mean squared error    :  13.01507
Sample standard deviation  :  3.135355

>
> #Principal Component Regression (PCR)
> library(pls)
> set.seed(2)
> pcr.fit <- pcr(fmla, data = qs, scale = TRUE, validation = "CV")
> summary(pcr.fit)
Data:   X dimension: 100 5
    Y dimension: 100 1
Fit method: svdpc
Number of components considered: 5

VALIDATION: RMSEP
Cross-validated using 10 random segments.
      (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps
CV          16.41    16.50    16.03    13.18    12.91    13.04
adjCV       16.41    16.48    16.01    13.09    12.88    13.00

TRAINING: % variance explained
                  1 comps  2 comps  3 comps  4 comps  5 comps
X                  37.131   61.290    79.22    94.29   100.00
Academic_Reputation    2.487    9.336    38.77    41.62    41.67
>
> #Partial Least Squares (PLS) regression
> set.seed(2)
> pls.fit <- plsr(fmla, data = qs, scale = TRUE, validation ="CV")
> summary(pls.fit)
Data:   X dimension: 100 5
    Y dimension: 100 1
Fit method: kernelpls
Number of components considered: 5

VALIDATION: RMSEP
Cross-validated using 10 random segments.
      (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps
CV          16.41    13.31    12.97    12.98    13.03    13.04
adjCV       16.41    13.26    12.94    12.95    13.00    13.00

TRAINING: % variance explained
                  1 comps  2 comps  3 comps  4 comps  5 comps
X                   23.00    54.33    77.90    91.42   100.00
Academic_Reputation    39.17    41.55    41.66    41.67    41.67
```