Project_2

W.M.C.C.M.Wijesingha S/18/836

2024-05-30

1. INTRODUCTION

Canonical Correlation Analysis (CCA) stands out as a powerful statistical technique designed specifically for identifying and quantifying relationships between two multivariate data sets. By uncovering underlying patterns and correlations, CCA provides valuable insights that can inform decision-making processes in fields ranging from economics and finance to psychology and biology. CCA allows us to summarize the relationship into lesser number of statistics while preserving the main facets of the relationships. This is another dimension reduction technique. The main goal of this study is to use Canonical Correlation Analysis to discover connections between two different sets of data. By using CCA, we want to find hidden relationships, patterns, and connections between variables in each dataset. This will help us better understand how the two datasets are related to each other. For that purpose here we use “College Freshmen data” dataset .

Hypothesis

In this study we test the hypothesis that canonical variate pairs are correlated or not (canonical correlations are equal or not equal to zero) .

H_0 : Canonical correlations are significant.
H_1 : Canonical correlations are not significant.

2. Methodology

2.1 Dataset

A researcher has collected data on three psychological variables (locus of control, self-concept, motivation), four academic variables(read, write, math, science)and gender of 600 college students. The researcher interest in identifying how the set of psychological variables relates to the academic variables and gender. Also the researcher is interested in how many dimensions are necessary to understand the relationship between the two set of variables.

2.2 Statistical Methods

The dataset used to this Canonical Correlation Analysis is consist of 8 variables and 600 observations. The variables are Locus of control, Self-concept, Motivation, Reading, Writing, Maths , Science and Gender is a zero-one indicator variable with the one indicating a female student. Canonical Correlation Analysis is a method for exploring the relationships between two multivariate sets of variables, all measured on the same individual. The multiple correlation measure a linear relationship between one Y and several x variables that we measure in regression analysis. Canonical Correlation Analysis determines a set of canonical variates, orthogonal linear combinations of the variables within each set that best explain the variability both within and between sets.

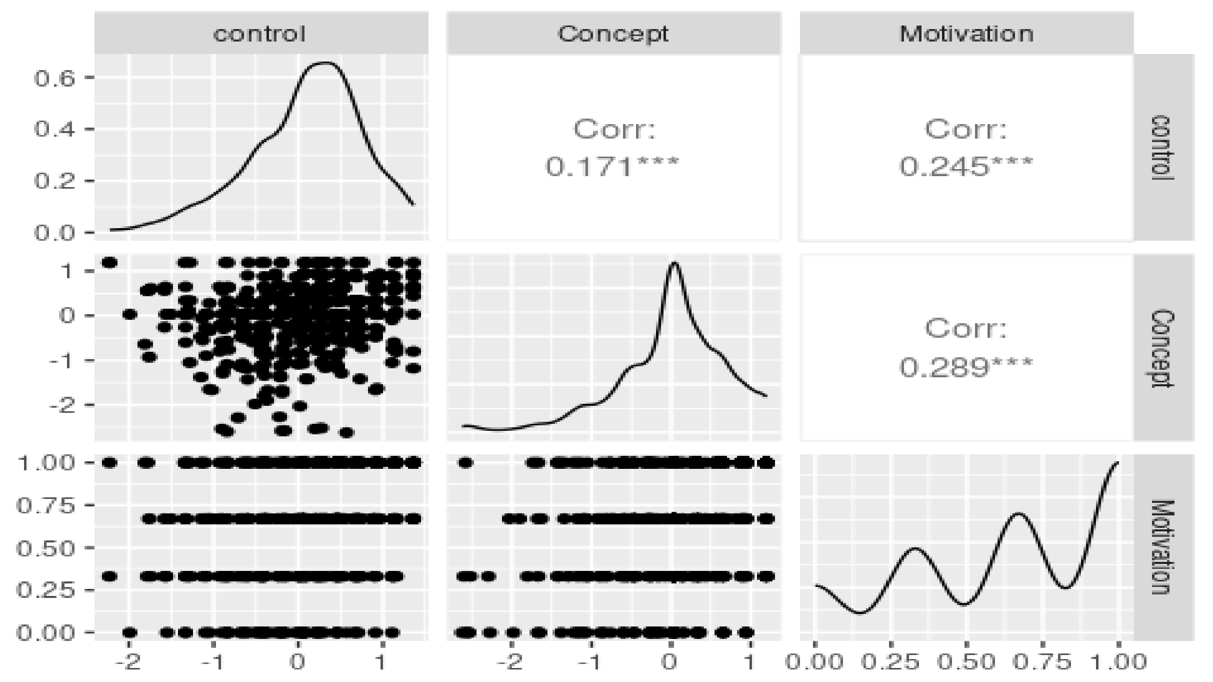
3. Results and Discussion

```
FreshmenData <- read_csv("../data/collegeFreshmenData.csv")

colnames(FreshmenData) <-c("control","Concept","Motivation","Read","Write","M
ath","Science","Sex")
summary(FreshmenData)

##     control             Concept             Motivation          Read     
##  Min.   :-2.23000   Min.   :-2.620000   Min.   :0.0000   Min.   :28.3  
##  1st Qu.:-0.37250   1st Qu.:-0.300000   1st Qu.:0.3300   1st Qu.:44.2  
##  Median : 0.21000   Median : 0.030000   Median :0.6700   Median :52.1  
##  Mean   : 0.09653   Mean   : 0.004917   Mean   :0.6608   Mean   :51.9  
##  3rd Qu.: 0.51000   3rd Qu.: 0.440000   3rd Qu.:1.0000   3rd Qu.:60.1  
##  Max.   : 1.36000   Max.   : 1.190000   Max.   :1.0000   Max.   :76.0  
##      Write            Math          Science           Sex       
##  Min.   :25.50   Min.   :31.80   Min.   :26.00   Min.   :0.000  
##  1st Qu.:44.30   1st Qu.:44.50   1st Qu.:44.40   1st Qu.:0.000  
##  Median :54.10   Median :51.30   Median :52.60   Median :1.000  
##  Mean   :52.38   Mean   :51.85   Mean   :51.76   Mean   :0.545  
##  3rd Qu.:59.90   3rd Qu.:58.38   3rd Qu.:58.65   3rd Qu.:1.000  
##  Max.   :67.10   Max.   :75.50   Max.   :74.20   Max.   :1.000  

xtabs(~Sex, data = FreshmenData)

## Sex
##   0   1 
## 273 327
```

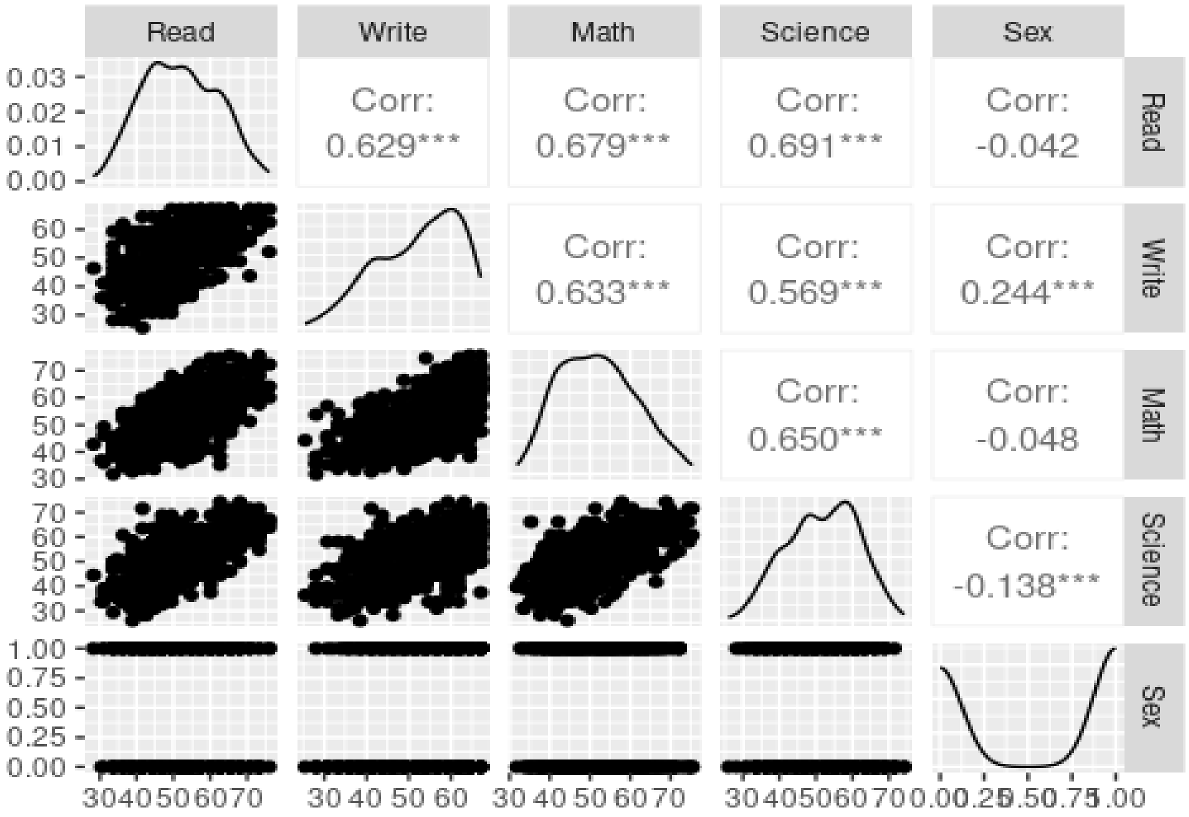
Canonical Correlation Analysis requires two sets of variables enclosed with a pair of parentheses. Psychological variables are specified as the first set of variables and the Academic variables with Gender as the second set.

```
psychological<- FreshmenData[, 1:3]
academic <- FreshmenData[, 4:8]
```

```
ggpairs(psychological)
```



```
ggpairs(academic)
```



correlations within and between the two sets of variables

```
matcor(psychological, academic)

## $Xcor
##              control   Concept Motivation
## control    1.0000000 0.1711878  0.2451323
## Concept    0.1711878 1.0000000  0.2885707
## Motivation 0.2451323 0.2885707  1.0000000
##
## $Ycor
##                Read     Write       Math    Science         Sex
## Read     1.00000000 0.6285909  0.6792757  0.6906929 -0.04174278
## Write    0.62859089 1.0000000  0.6326664  0.5691498  0.24433183
## Math     0.67927568 0.6326664  1.0000000  0.6495261 -0.04821830
## Science  0.69069291 0.5691498  0.6495261  1.0000000 -0.13818587
## Sex     -0.04174278 0.2443318 -0.0482183 -0.1381859  1.00000000
##
## $XYcor
##              control     Concept Motivation        Read       Write        M
```

```
ath
## control    1.0000000  0.17118778 0.24513227  0.37356505 0.35887684  0.3372
690
## Concept    0.1711878  1.00000000 0.28857075  0.06065584 0.01944856  0.0535
977
## Motivation 0.2451323  0.28857075 1.00000000  0.21060992 0.25424818  0.1950
135
## Read       0.3735650  0.06065584 0.21060992  1.00000000 0.62859089  0.6792
757
## Write      0.3588768  0.01944856 0.25424818  0.62859089 1.00000000  0.6326
664
## Math       0.3372690  0.05359770 0.19501347  0.67927568 0.63266640  1.0000
000
## Science    0.3246269  0.06982633 0.11566948  0.69069291 0.56914983  0.6495
261
## Sex        0.1134108 -0.12595132 0.09810277 -0.04174278 0.24433183 -0.0482
183
##                Science         Sex
## control     0.32462694  0.11341075
## Concept     0.06982633 -0.12595132
## Motivation  0.11566948  0.09810277
## Read        0.69069291 -0.04174278
## Write       0.56914983  0.24433183
## Math        0.64952612 -0.04821830
## Science     1.00000000 -0.13818587
## Sex        -0.13818587  1.00000000
```

Canonical Correlations

```
canoncor1<- cc(psychological,academic)
canoncor1$cor
```

```
## [1] 0.4640861 0.1675092 0.1039911
```

Raw Canonical Coefficients

```
canoncor1[3:4]
```

```
## $xcoef
##                  [,1]       [,2]       [,3]
## control    -1.2538339 -0.6214776 -0.6616896
## Concept     0.3513499 -1.1876866  0.8267210
## Motivation -1.2624204  2.0272641  2.0002283
##
## $ycoef
##                 [,1]         [,2]         [,3]
## Read    -0.044620600 -0.004910024  0.021380576
## Write   -0.035877112  0.042071478  0.091307329
## Math    -0.023417185  0.004229478  0.009398182
## Science -0.005025152 -0.085162184 -0.109835014
## Sex     -0.632119234  1.084642326 -1.794647036
```

For the variable Read, a one unit increase in reading leads to a 0.446 decrease in the first canonical variate of set 2 when all of the other variables are held constant. similarly when the other predictors held constant, for the variable Write, a one unit increase in writing leads to a 0.0358 decrease in the first canonical variate of set 2 for the variable Math, a one unit increase in writing leads to a 0.0234 decrease in the first canonical variate of set 2 for the variable Science, a one unit increase in writing leads to a 0.005 decrease in the first canonical variate of set 2 for the variable Sex, being female leads to a 0.632 decrease in the first canonical variate of set 2.

Canonical loadings- correlations between variables and the canonical variates

```
canoncor2 <- comput(psychological, academic, canoncor1)
canoncor2[3:6]
```

```
## $corr.X.xscores
##                   [,1]       [,2]       [,3]
## control    -0.90404631 -0.3896883 -0.1756227
## Concept    -0.02084327 -0.7087386  0.7051632
## Motivation -0.56715106  0.3508882  0.7451289
## 
## $corr.Y.xscores
##               [,1]        [,2]        [,3]
## Read    -0.3900402 -0.06010654  0.01407661
## Write   -0.4067914  0.01086075  0.02647207
## Math    -0.3545378 -0.04990916  0.01536585
## Science -0.3055607 -0.11336980 -0.02395489
## Sex     -0.1689796  0.12645737 -0.05650916
## 
## $corr.X.yscores
##                    [,1]        [,2]        [,3]
## control    -0.419555307 -0.06527635 -0.01826320
## Concept    -0.009673069 -0.11872021  0.07333073
## Motivation -0.263206910  0.05877699  0.07748681
## 
## $corr.Y.yscores
##               [,1]        [,2]       [,3]
## Read    -0.8404480 -0.35882541  0.1353635
## Write   -0.8765429  0.06483674  0.2545608
## Math    -0.7639483 -0.29794884  0.1477611
## Science -0.6584139 -0.67679761 -0.2303551
## Sex     -0.3641127  0.75492811 -0.5434036
```

Tests of the Canonical Dimensions

```
library(CCP)
rho <- canoncor1$cor

## Define number of of observations
n<-dim(psychological)[1]

## Number of variables in the first set
p <- length(psychological)

## Number of variables in the second set
q <- length(academic)
```

Calculate p-value using the F-approximations od different test statistics

```
p.asym(rho,n,p,q,tstat ="Wilks")

## Wilks' Lambda, using F-approximation (Rao's F):
##                stat    approx df1      df2     p.value
## 1 to 3:  0.7543611 11.715733  15 1634.653 0.000000000
## 2 to 3:  0.9614300  2.944459   8 1186.000 0.002905057
## 3 to 3:  0.9891858  2.164612   3  594.000 0.091092180

p.asym(rho,n,p,q, tstat = "Hotelling")

##  Hotelling-Lawley Trace, using F-approximation:
##                 stat    approx df1  df2     p.value
## 1 to 3:  0.31429738 12.376333  15 1772 0.000000000
## 2 to 3:  0.03980175  2.948647   8 1778 0.002806614
## 3 to 3:  0.01093238  2.167041   3 1784 0.090013176

p.asym(rho,n,p,q,tstat = "Pillai")

##  Pillai-Bartlett Trace, using F-approximation:
##                 stat    approx df1  df2     p.value
## 1 to 3:  0.25424936 11.000571  15 1782 0.000000000
## 2 to 3:  0.03887348  2.934093   8 1788 0.002932565
## 3 to 3:  0.01081416  2.163421   3 1794 0.090440474

p.asym(rho,n,p,q,tstat = "Roy")

##  Roy's Largest Root, using F-approximation:
##                stat   approx df1 df2 p.value
## 1 to 1:  0.2153759 32.61008   5 594       0
##
##  F statistic for Roy's Greatest Root is an upper bound.
```

The first test of the Canonical dimensions tests whether all three dimensions are significant(they are, F=11.72), the next test tests whether dimensions 2 and 3 combined are significant(they are,F=2.94). Finally, the last test tests whether dimension 3, by itself, is significant(it is not) Therefore considering all the above tests, dimension 1 and 2 must each be significant while dimension 3 is not.
When the variables in the model have very different standard deviations, the standardized coefficients allow for easier comparison among the variables.

Standardized psychological Canonical coefficients

```
s1 <- diag(sqrt(diag(cov(psychological))))
s1 %*%
canoncor1$xcoef

##            [,1]       [,2]       [,3]
## [1,] -0.8404196 -0.4165639 -0.4435172
## [2,]  0.2478818 -0.8379278  0.5832620
## [3,] -0.4326685  0.6948029  0.6855370
```

Standardized academic Canonical coefficients

```
s2<- diag(sqrt(diag(cov(academic))))
s2 %*%
canoncor1$ycoef

##             [,1]        [,2]        [,3]
## [1,] -0.45080116 -0.04960589  0.21600760
## [2,] -0.34895712  0.40920634  0.88809662
## [3,] -0.22046662  0.03981942  0.08848141
## [4,] -0.04877502 -0.82659938 -1.06607828
## [5,] -0.31503962  0.54057096 -0.89442764
```

The variable Read, a one standard deviation increase in reading leads to a 0.45 standard deviation decrease in the score on the first canonical variate for set 2 when the other variables in the model are held constant.

4. Conclusion and recommendation

Tests of Canonical Dimensions

```
## Dimension Canonical_corr Mult_F df1 df2 p_value
## 1 1 0.4641 11.72 15 1634.7 0.0000
## 2 2 0.1675 2.94 8 1186.0 0.0029
## 3 3 0.1040 2.16 3 594.0 0.0911
```

Tests of dimensionality for the canonical correlation analysis, as shown in above table, indicate that two of the three canonical dimensions are statistically significant at the 0.05significant level. Dimension 1 had a canonical correlation of 0.4641 between the sets of variables, while for dimension 2 the canonical correlation was much lower at 0.1675.

Standardized Canonical Coefficients

```
## Psychological_variables Dim_1 Dim_2
## 1 locus of control -0.8404 -0.4166
## 2 self-concept 0.2479 -0.8379
## 3 motivation -0.4327 0.6948
## Academic_variables Dim1 Dim2
## 1 Reading -0.4508 -0.0496
## 2 Writing 0.2479 0.4092
## 3 Maths -0.2205 0.0398
## 4 Science -0.0488 -0.8266
## 5 Gender -0.3150 0.5406
```

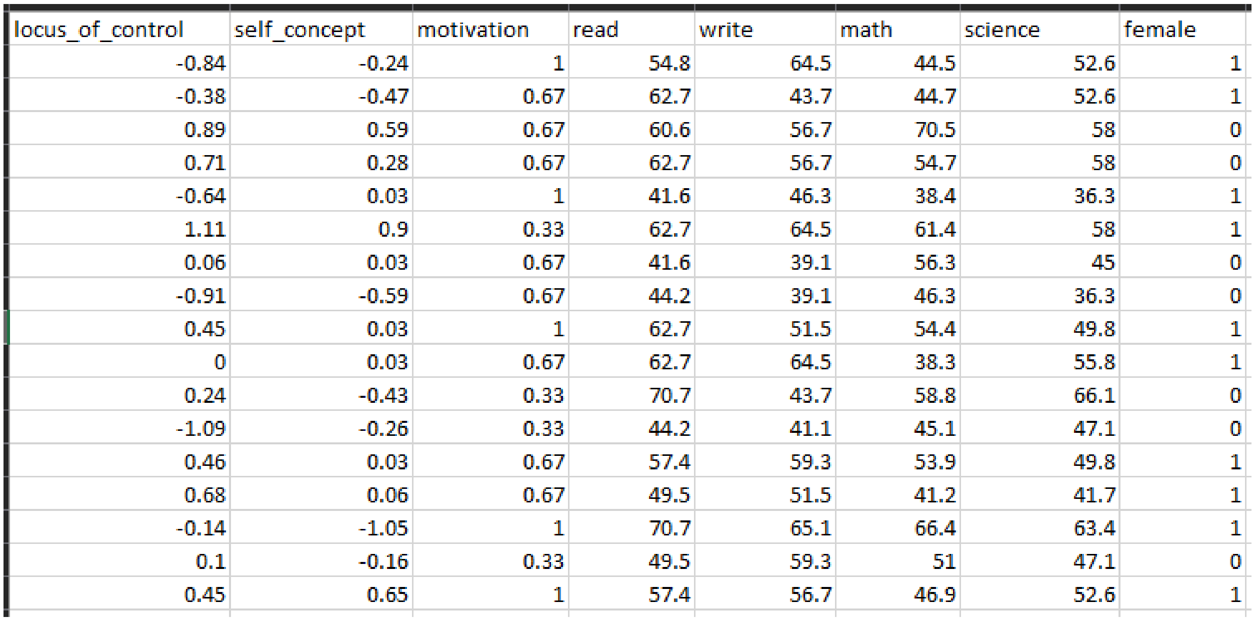
The standardized canonical coefficients for the first two dimensions across both sets of variables are given in the above table. For the psychological variables, the first canonical dimension is most strongly influenced by locus of control (-0.8404) and for the second canonical dimension is most influences by self-concept (-0.8379) and motivation (0.6948). For academic variables with Gender, the first canonical dimension was comprised of Reading (-0.4508), writing (-0.3489) and gender (-0.3150) and for the second canonical dimension writing (0.4092), science (-0.8266) and gender (0.5406) were the dominating variables.

5. References

Thurstone, L. L. (1947). Multiple factor analysis. University of Chicago Press.
Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate data analysis (8th ed.). Cengage Learning.

6. Appendices

6.1 Dataset



locus_of_control	self_concept	motivation	read	write	math	science	female
-0.84	-0.24	1	54.8	64.5	44.5	52.6	1
-0.38	-0.47	0.67	62.7	43.7	44.7	52.6	1
0.89	0.59	0.67	60.6	56.7	70.5	58	0
0.71	0.28	0.67	62.7	56.7	54.7	58	0
-0.64	0.03	1	41.6	46.3	38.4	36.3	1
1.11	0.9	0.33	62.7	64.5	61.4	58	1
0.06	0.03	0.67	41.6	39.1	56.3	45	0
-0.91	-0.59	0.67	44.2	39.1	46.3	36.3	0
0.45	0.03	1	62.7	51.5	54.4	49.8	1
0	0.03	0.67	62.7	64.5	38.3	55.8	1
0.24	-0.43	0.33	70.7	43.7	58.8	66.1	0
-1.09	-0.26	0.33	44.2	41.1	45.1	47.1	0
0.46	0.03	0.67	57.4	59.3	53.9	49.8	1
0.68	0.06	0.67	49.5	51.5	41.2	41.7	1
-0.14	-1.05	1	70.7	65.1	66.4	63.4	1
0.1	-0.16	0.33	49.5	59.3	51	47.1	0
0.45	0.65	1	57.4	56.7	46.9	52.6	1

6.2 R codes

❖
```
library(tidyverse)

library(GGally)

library(CCP)

library(CCA)
```

❖
```
FreshmenData <- read_csv("../data/collegeFreshmenData.csv")

colnames(FreshmenData) <-
c("control","Concept","Motivation","Read","Write","Math","Science","Sex")

summary(FreshmenData)
```

❖
```
xtabs(~Sex, data = FreshmenData)
```

❖
```
psychological<- FreshmenData[, 1:3]

academic <- FreshmenData[, 4:8]

ggpairs(psychological)
```

-
  ```
  ggpairs(academic)
  matcor(psychological, academic)
  ```

-
  ```
  canoncor1<- cc(psychological,academic)
  canoncor1$cor
  canoncor1[3:4]
  canoncor2 <- comput(psychological, academic, canoncor1)
  canoncor2[3:6]
  ```

-
  ```
  rho <- canoncor1$cor
  ```

-
  ```
  ## Define number of of observations
  n<-dim(psychological)[1]
  ## Number of variables in the first set
  p <- length(psychological)
  ## Number of variables in the second set
  q <- length(academic)
  ```

-
  ```
  p.asym(rho,n,p,q,tstat ="Wilks")
  p.asym(rho,n,p,q, tstat = "Hotelling")
  p.asym(rho,n,p,q,tstat = "Pillai")
  p.asym(rho,n,p,q,tstat = "Roy")
  ```

-
  ```
  s1 <- diag(sqrt(diag(cov(psychological))))
  s1 %*%
  canoncor1$xcoef
  ```

-
  ```
  s2<- diag(sqrt(diag(cov(academic))))
  s2 %*%
  canoncor1$ycoef
  ```