# Canonical Correlation Analysis

Humaun Farid Sohag

2025-08-13

## #Step 1. Load dataset

```
mm <- read.csv("mmreg.csv")

# Rename columns for clarity
colnames(mm) <- c("Control", "Concept", "Motivation", "Read", "Write", "Math", "Science", "Sex")
colnames(mm)
```

```
## [1] "Control"    "Concept"    "Motivation" "Read"       "Write"
## [6] "Math"       "Science"    "Sex"
```

```
head(mm)
```

```
##   Control Concept Motivation Read Write Math Science Sex
## 1   -0.84   -0.24       1.00 54.8  64.5 44.5    52.6   1
## 2   -0.38   -0.47       0.67 62.7  43.7 44.7    52.6   1
## 3    0.89    0.59       0.67 60.6  56.7 70.5    58.0   0
## 4    0.71    0.28       0.67 62.7  56.7 54.7    58.0   0
## 5   -0.64    0.03       1.00 41.6  46.3 38.4    36.3   1
## 6    1.11    0.90       0.33 62.7  64.5 61.4    58.0   1
```

## #Step 2. Descriptive statistics

```
summary(mm)
```

```
##     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
```

```r
xtabs(~Sex, data = mm)  # frequency table for Sex
```

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

## #Step 3. Load required packages

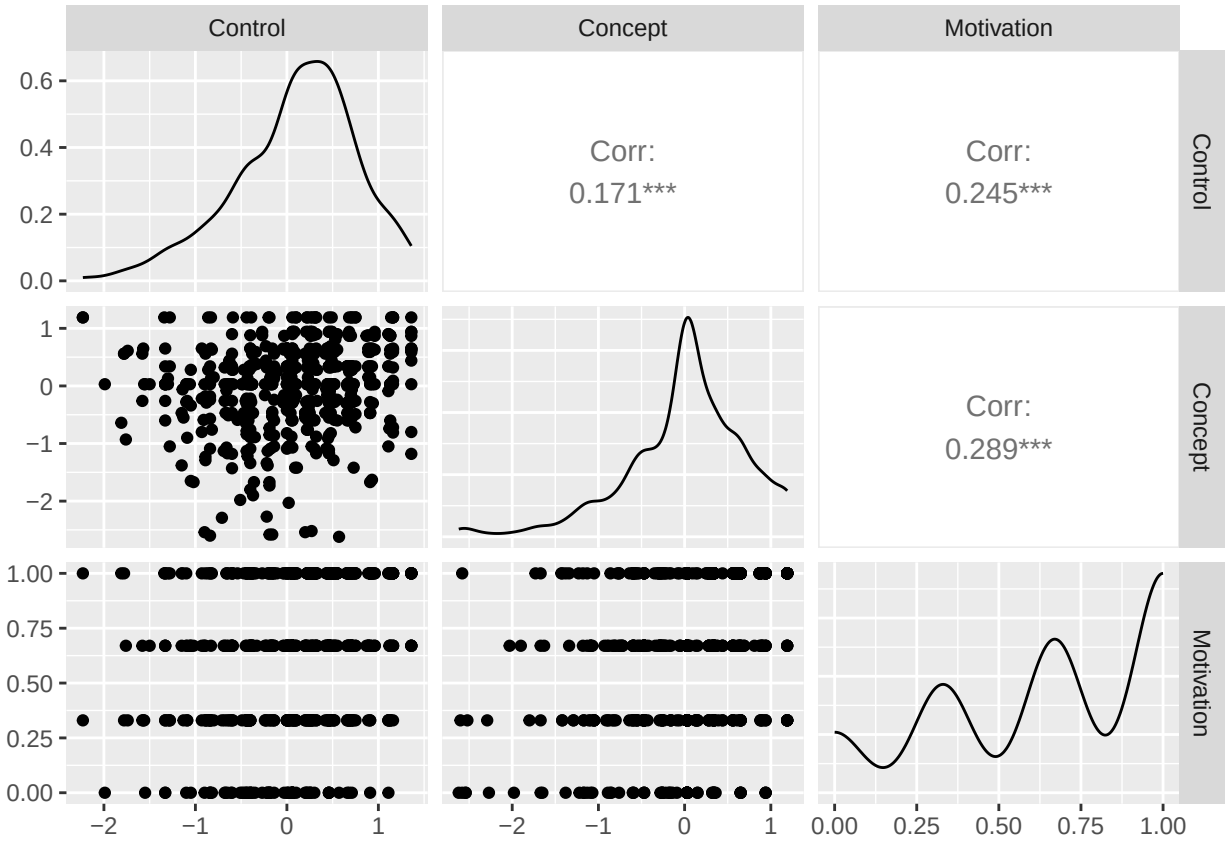
```r
library(ggplot2)
library(GGally)   # for ggpairs
library(CCA)      # Canonical Correlation Analysis
library(CCP)      # Significance testing of canonical correlations
```

## #Step 4. Split dataset into psychological and academic variables

```r
psych <- mm[, 1:3]  # locus_of_control, self_concept, motivation
acad  <- mm[, 4:8]  # read, write, math, science, sex
```

## #Step 5. Exploratory data analysis

```r
ggpairs(psych)  # pairwise plots for psychological variables
```




## #Step 6. Compute correlation matrix between two sets

```
matcor(psych, acad)
```

```
## $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       Math
## Control    1.0000000  0.17118778 0.24513227  0.37356505 0.35887684  0.3372690
## Concept    0.1711878  1.00000000 0.28857075  0.06065584 0.01944856  0.0535977
## Motivation 0.2451323  0.28857075 1.00000000  0.21060992 0.25424818  0.1950135
```

```
## Read       0.3735650  0.06065584 0.21060992  1.00000000 0.62859089  0.6792757
## Write      0.3588768  0.01944856 0.25424818  0.62859089 1.00000000  0.6326664
## Math       0.3372690  0.05359770 0.19501347  0.67927568 0.63266640  1.0000000
## Science    0.3246269  0.06982633 0.11566948  0.69069291 0.56914983  0.6495261
## Sex        0.1134108 -0.12595132 0.09810277 -0.04174278 0.24433183 -0.0482183
##                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
```

## #Step 7. Perform canonical correlation analysis

```
cc1 <- cc(psych, acad)

# Display canonical correlations
cc1$cor
```

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

## #Step 8. Compute canonical loadings

```
cc2 <- comput(psych, acad, cc1)

# Display canonical loadings (X & Y scores)
cc2[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
```

## #Step 9. Tests of canonical dimensions

```
rho <- cc1$cor
n   <- nrow(psych)   # number of observations
p   <- ncol(psych)   # number of variables in set 1
q   <- ncol(acad)    # number of variables in set 2

# Calculate p-values using different F-approximations
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.
```

## #Step 10. Standardized canonical coefficients

```
# Psychological variables
s1 <- diag(sqrt(diag(cov(psych))))
std_xcoef <- s1 %*% cc1$xcoef
std_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
```

```
# Academic variables
s2 <- diag(sqrt(diag(cov(acad))))
std_ycoef <- s2 %*% cc1$ycoef
std_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
```