REPORT

ON

CANONICAL CORRELATION ANALYSIS

BY:

Thivajini Ravichandran

S/18/824

1. Introduction

Canonical correlation analysis is a method for exploring the relationship between two multivariate sets of variables, all measured on the same individual. It 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. These are the main objectives of Canonical Correlation Analysis.

In here, we test the hypothesis that canonical variate pairs are correlated or not (canonical correlations are equal or not equal to zero) at 5% significance level.

I use "mmreg" dataset and investigating the associations between psychological measures and academic achievement measures in this dataset using canonical correlation analysis.

2. Methodology

Description of the dataset

I have a data file, **mmreg.data**, with 600 observations on eight variables.

The psychological variable are **locus of control**, **self concept** and **motivation**.

The academic variables are standardized tests in reading (**read**), writing (**write**), math (**math**) and science (**science**).

Additionally, the variable **female** is a zero-one indicator variable with the one indicating a female student.

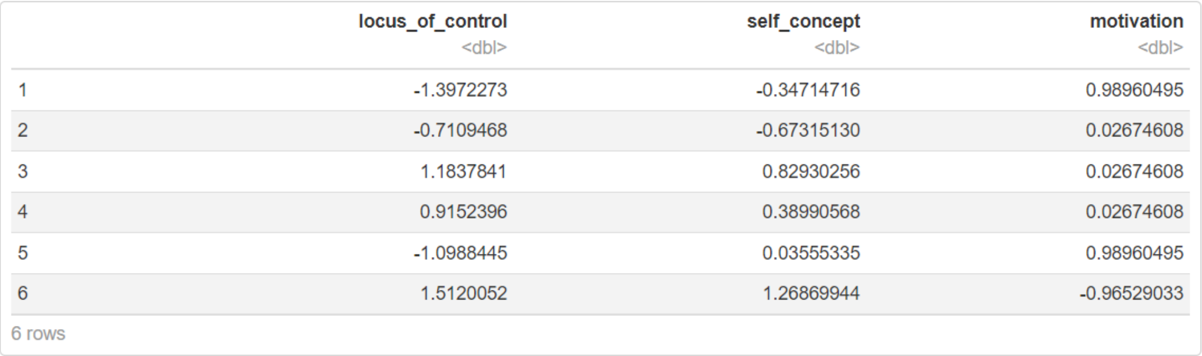
Use Methods:

- Canonical Correlation Analysis.

3. Results and Discussion

Split the dataset into 2 sets of variable. We specify our psychological variables as the first set of variables and our academic variables plus gender as the second set. For convenience, the variables in the first set are called "u" variables and the variables in the second set are called "v" variables. There are p = 3 variables in the first group relating to psychology and q = 5 variables in the second group relating to academic.

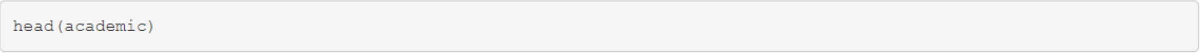
- **Psychological variables:**



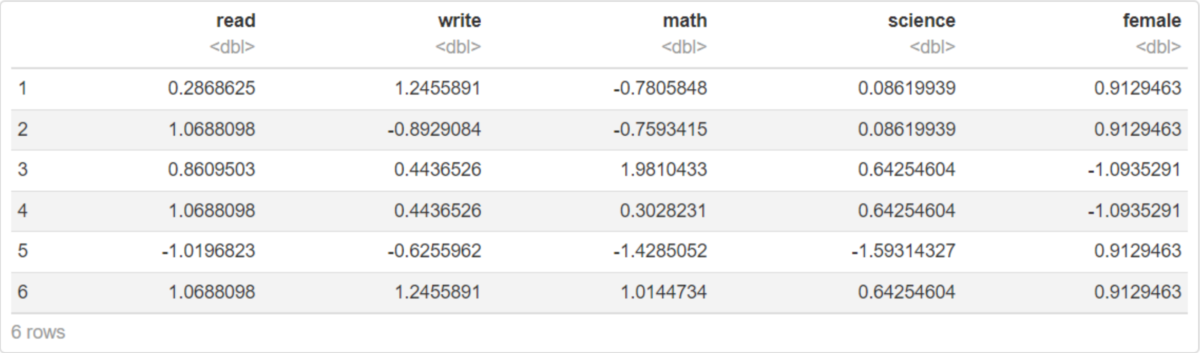
	locus_of_control <dbl>	self_concept <dbl>	motivation <dbl>
1	-1.3972273	-0.34714716	0.98960495
2	-0.7109468	-0.67315130	0.02674608
3	1.1837841	0.82930256	0.02674608
4	0.9152396	0.38990568	0.02674608
5	-1.0988445	0.03555335	0.98960495
6	1.5120052	1.26869944	-0.96529033

6 rows

- **Academic variables:**



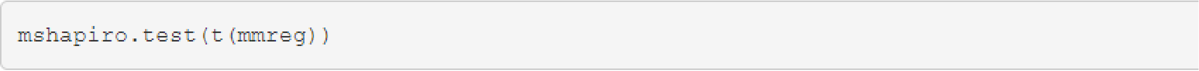
```
head(academic)
```



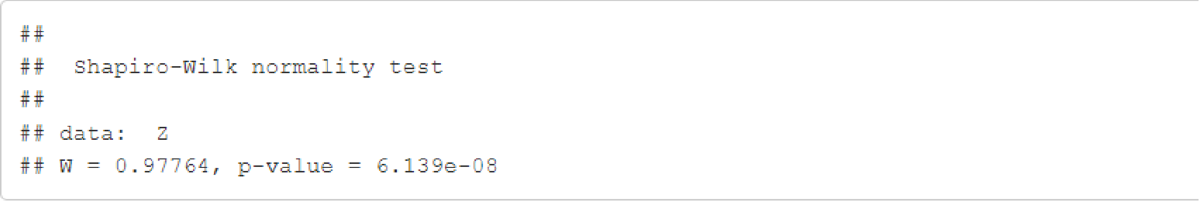
	read <dbl>	write <dbl>	math <dbl>	science <dbl>	female <dbl>
1	0.2868625	1.2455891	-0.7805848	0.08619939	0.9129463
2	1.0688098	-0.8929084	-0.7593415	0.08619939	0.9129463
3	0.8609503	0.4436526	1.9810433	0.64254604	-1.0935291
4	1.0688098	0.4436526	0.3028231	0.64254604	-1.0935291
5	-1.0196823	-0.6255962	-1.4285052	-1.59314327	0.9129463
6	1.0688098	1.2455891	1.0144734	0.64254604	0.9129463

6 rows

- **Testing Normality**



```
mshapiro.test(t(mmreg))
```

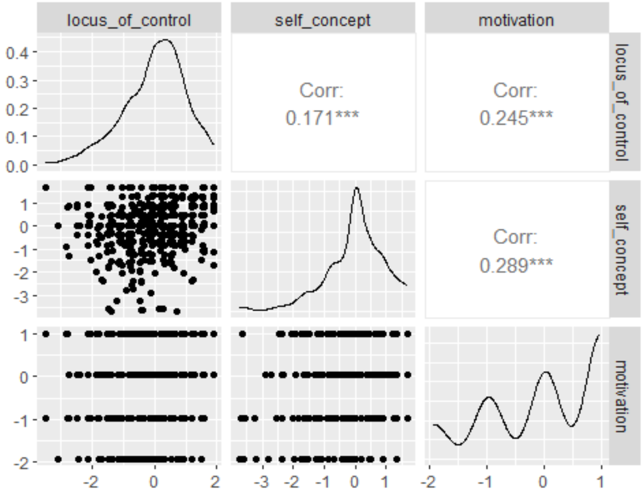


```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.97764, p-value = 6.139e-08
```

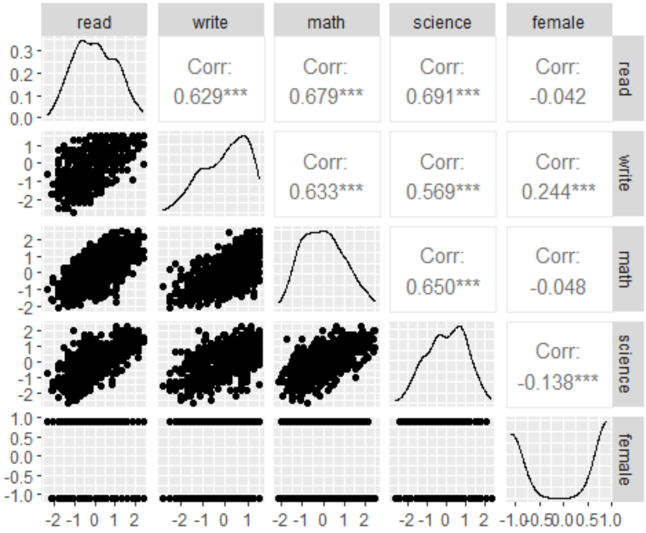
Before performing the canonical analysis, a multivariate normality test should be performed as the first assumption. Here we did a multivariate Shapiro-Wilk normality test, but since our p value is less than 0.05, we see that our data set is not normally distributed.

- **Check whether there is a correlation between variables in the two sets of variables.**

Psychological variables set



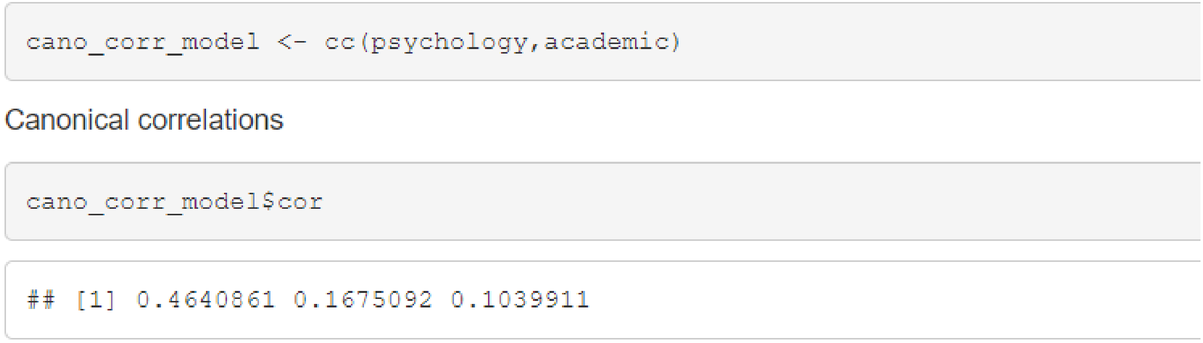
Academic variables set



As you can see there is significant correlation among variables in the academic set of variables and weak correlation among psychological variables.

- **Canonical correlation analysis**

Fit the canonical correlation model and get canonical correlations. We can conclude that there are 3 (equal to number of variables in set “psychology”; small set) canonical correlations in this model.



```
cano_corr_model <- cc(psychology,academic)
```

Canonical correlations

```
cano_corr_model$cor
```

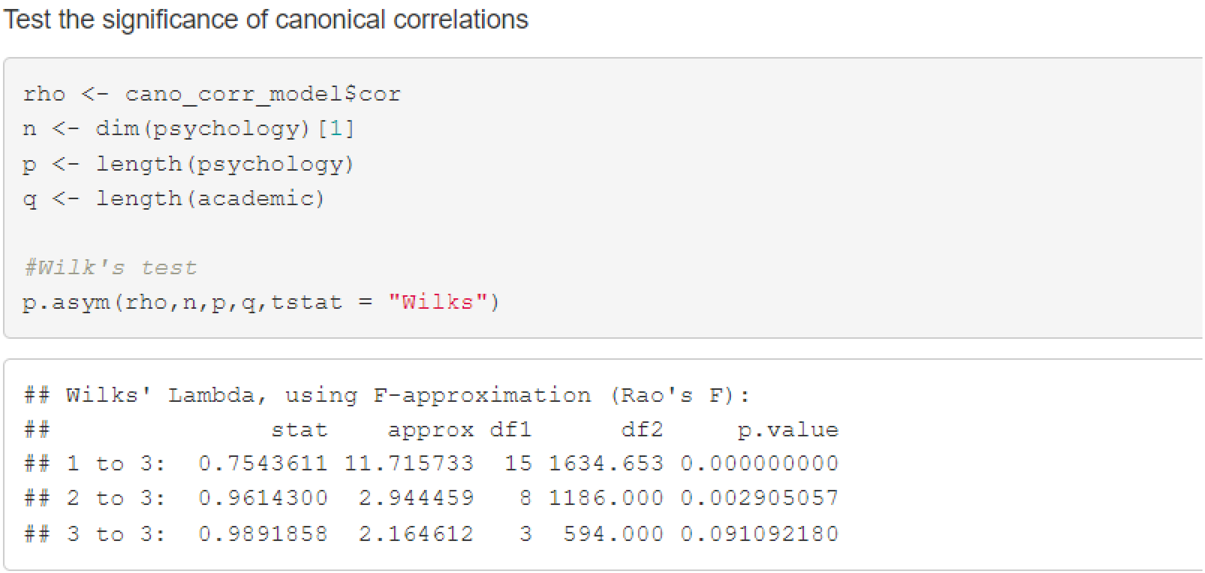
```
## [1] 0.4640861 0.1675092 0.1039911
```

Our coefficients of canonical correlation are as seen above. Its significance must be tested before interpreting these coefficients. We will do the meaningfulness test with the Wilk’s test. In general, the number of canonical dimensions is equal to the number of variables in the smaller set. However, the number of significant dimensions may be even smaller. Canonical dimensions, also known as canonical variates, are latent variables that are analogous to factors obtained in factor analysis. For this particular model there are three canonical dimensions of which only the first two are statistically significant. For statistical test, we use R package **"CCP"**.

- **Test for relationship between canonical variate pairs**

H_0: Regression coefficients (except for the intercepts) are all equal to zero.

which means first set of variables is independent from the second set of variables.



Test the significance of canonical correlations

```
rho <- cano_corr_model$cor
n <- dim(psychology)[1]
p <- length(psychology)
q <- length(academic)

#Wilk's test
p.asym(rho,n,p,q,tstat = "Wilks")
```

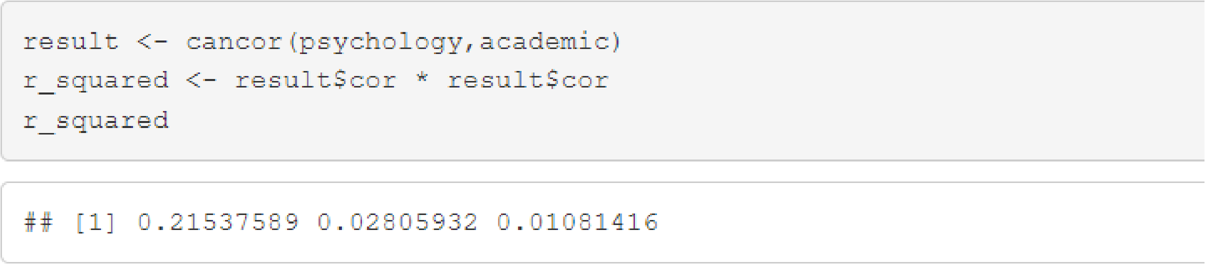
```
## Wilks' Lambda, using F-approximation (Rao's F):
##                 stat    approx df1      df2     p.value
## 1 to 3:  0.7543611 11.715733  15 1634.653 0.000000000
## 2 to 3:  0.9614300  2.944459   8 1186.000 0.002905057
## 3 to 3:  0.9891858  2.164612   3  594.000 0.091092180
```

As shown in above, wilk’s lambda Λ = 0.75, F =11.72, p – value < 0.05.Here, we reject the null hypothesis and conclude that the two sets of variables are dependent.

Next consider second set, wilk’s lambda Λ = 0.96, F =2.94, p – value < 0.05.Then, we reject the null hypothesis and conclude that second canonical variate pair is correlated

Then wilk’s lambda Λ = 0.99, F =2.16, p – value > 0.05. We do not reject the null hypothesis and third canonical variate pair is not correlated and the analysis may be stopped.

- **Estimates of canonical correlation**



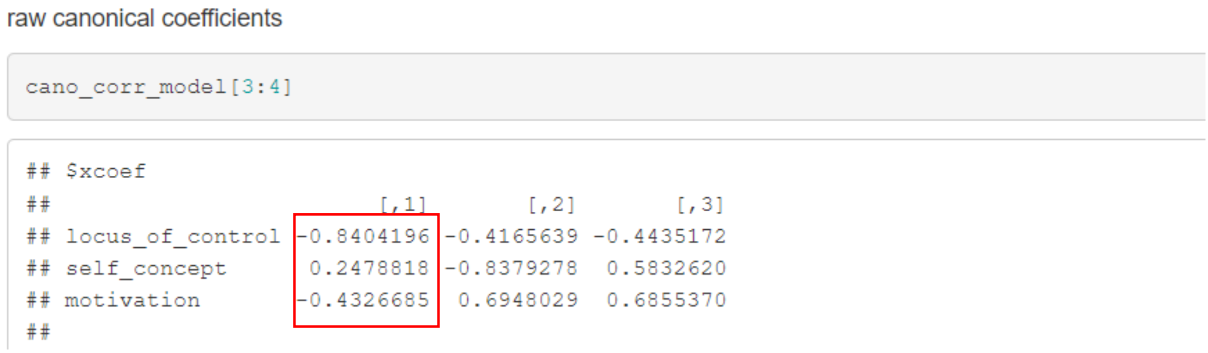
```
result <- cancor(psychology,academic)
r_squared <- result$cor * result$cor
r_squared
```

```
## [1] 0.21537589 0.02805932 0.01081416
```

We can interpret these results as follows. Here 21.54% of the variation in U_1 is explained by variation in V_1, 2.81% of the variation in U_2 is explained by variation in V_2 and 1.08% of the variation in U_3 is explained by variation in V_3. Since first canonical correlation is higher than others, it tells us first canonical correlation is very important.

- **Canonical coefficients**

The estimated canonical coefficients (a_{ij}) for the psychological variables



raw canonical coefficients

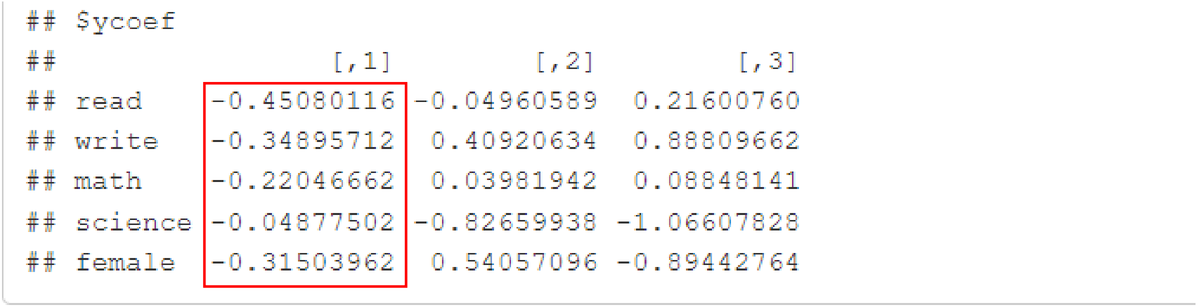
```
cano_corr_model[3:4]
```

```
## $xcoef
##                        [,1]       [,2]       [,3]
## locus_of_control -0.8404196 -0.4165639 -0.4435172
## self_concept      0.2478818 -0.8379278  0.5832620
## motivation       -0.4326685  0.6948029  0.6855370
##
```

The first canonical variable for psychology

$$U_1 = -(0.8404) * X_{locus\_of\_control} + (0.2479) * X_{self\_concept} - (0.4327) * X_{motivation}$$

The estimated canonical coefficients (b_{ij}) for the academic variables



```
## $ycoef
##                [,1]        [,2]        [,3]
## read    -0.45080116 -0.04960589  0.21600760
## write   -0.34895712  0.40920634  0.88809662
## math    -0.22046662  0.03981942  0.08848141
## science -0.04877502 -0.82659938 -1.06607828
## female  -0.31503962  0.54057096 -0.89442764
```

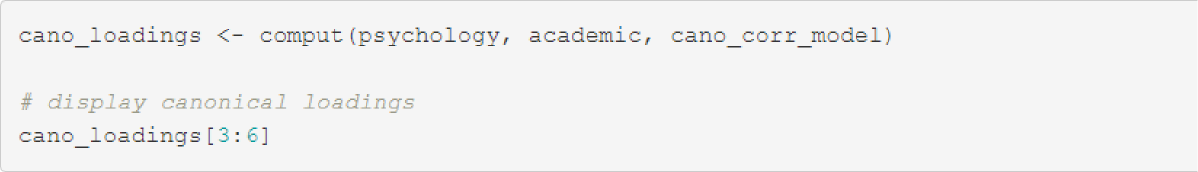
The first canonical variable for academic

$$V_1 = -(0.4508) * Y_{read} - (0.34895) * Y_{write} - (0.2205) * Y_{math} - (0.0488) * Y_{science} - (0.3150) * Y_{female}$$

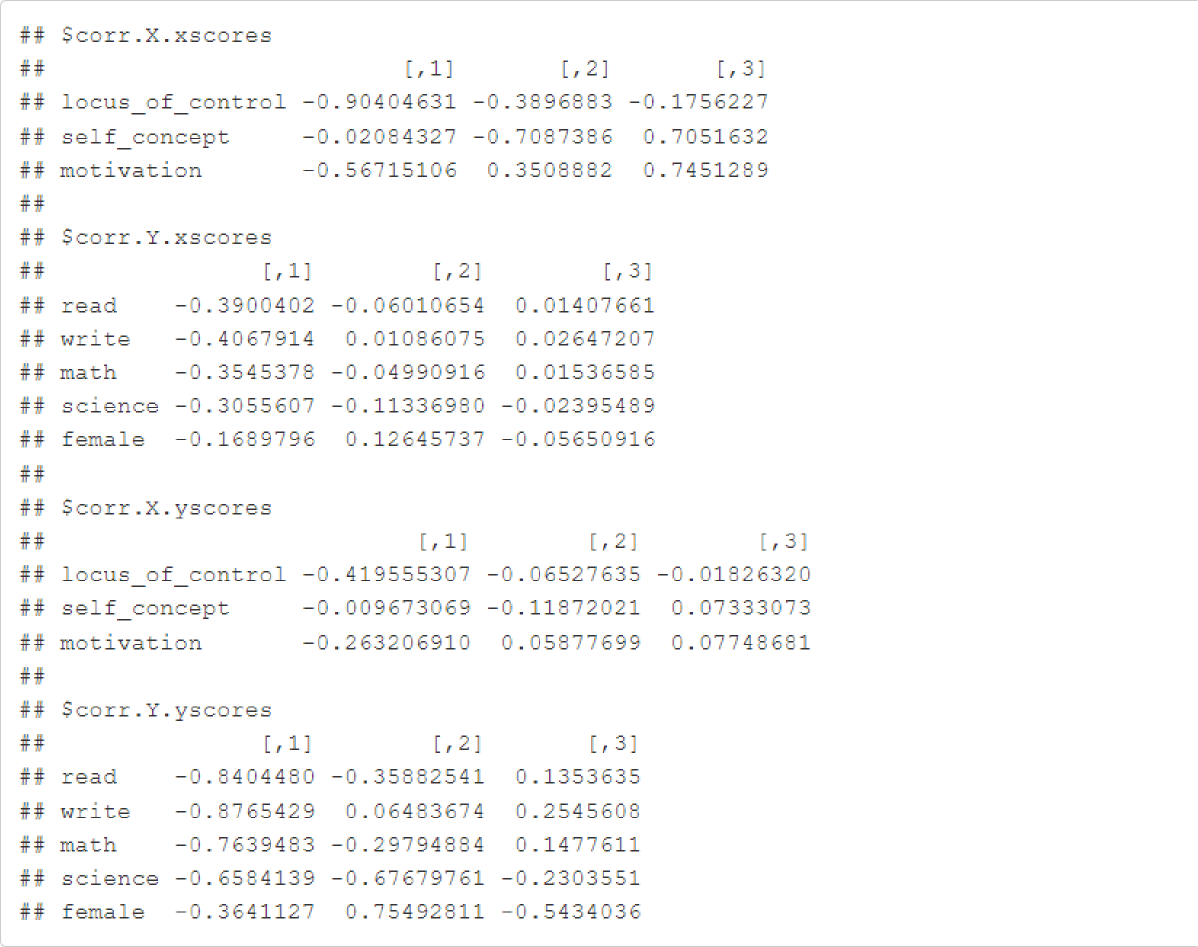
- **Interpret each component**

To interpret each component, we must compute the correlations between each variable and the corresponding canonical variate.

We'll use **comput** function to compute the loadings of the variables on the canonical dimensions (variates). These loadings are correlations between variables and the canonical variates.



```
cano_loadings <- comput(psychology, academic, cano_corr_model)

# display canonical loadings
cano_loadings[3:6]
```

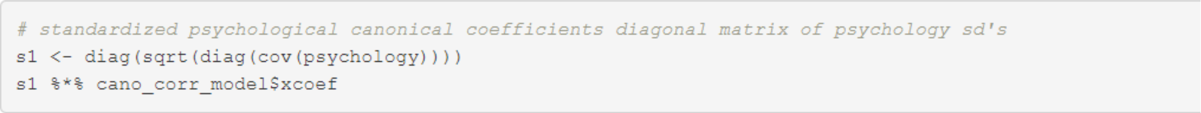


```
## $corr.X.xscores
##                         [,1]       [,2]       [,3]
## locus_of_control -0.90404631 -0.3896883 -0.1756227
## self_concept     -0.02084327 -0.7087386  0.7051632
## motivation       -0.56715106  0.3508882  0.7451289
##
## $corr.Y.xscores
##               [,1]        [,2]        [,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
## female  -0.1689796  0.12645737 -0.05650916
##
## $corr.X.yscores
##                          [,1]        [,2]        [,3]
## locus_of_control -0.419555307 -0.06527635 -0.01826320
## self_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
## female  -0.3641127  0.75492811 -0.5434036
```

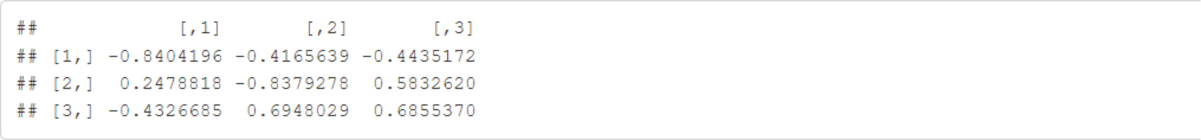
The above correlations are between observed variables and canonical variables which are known as the canonical loadings. These canonical variates are actually a type of latent variable.

- **Standardized Canonical Weights**

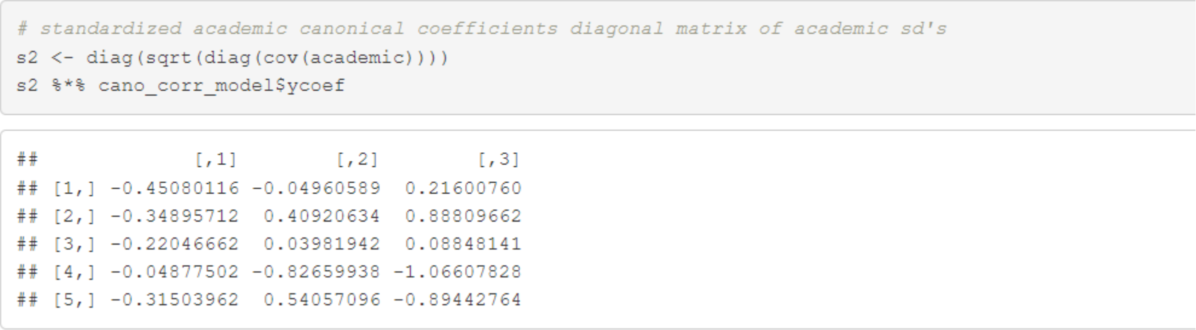
Since variables in the data set contain different standard deviations, we will perform standardization and examine their canonical coefficients on a variable basis in order to increase their interpretability.



```
# standardized psychological canonical coefficients diagonal matrix of psychology sd's
s1 <- diag(sqrt(diag(cov(psychology))))
s1 %*% cano_corr_model$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 diagonal matrix of academic sd's
s2 <- diag(sqrt(diag(cov(academic))))
s2 %*% cano_corr_model$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 standardized canonical coefficients are interpreted in a manner analogous to interpreting standardized regression coefficients. For example, consider 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 two when the other variables in the model are held constant. Similarly, we can interpret other coefficients as well.

4. Conclusion and Recommendation

From the analysis we can conclude that instead of 15 pairwise scatterplots comparison we can do it from only 3 canonical variate pairs. I proved it by "Wilki's lambda" test and also we have used same test to determine the significance of canonical variate pairs. According to these results we obtain the following conclusions.

1. There are two significant canonical covariate pairs which can be used for our analysis.

2. According to squared canonical values 21.54% of the variation in first canonical variable of psychological set is explained by that of first canonical variable of academic set.

5. References

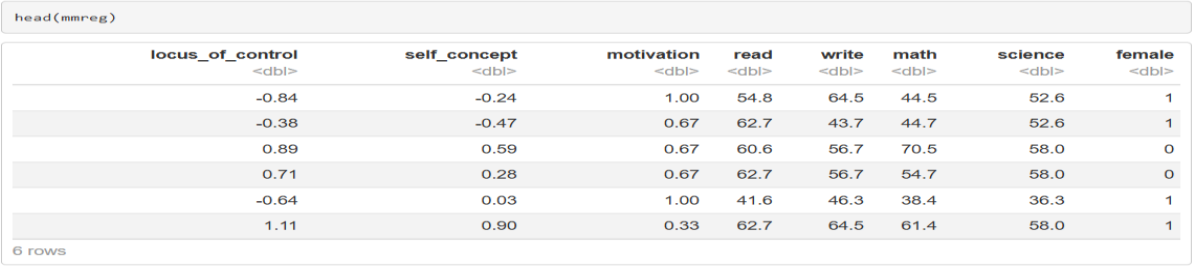
References

https://stats.oarc.ucla.edu

https://www.geeksforgeeks.org/canonical-correlation-analysis-cca-using-sklearn

6. Appendices

- Part of the dataset

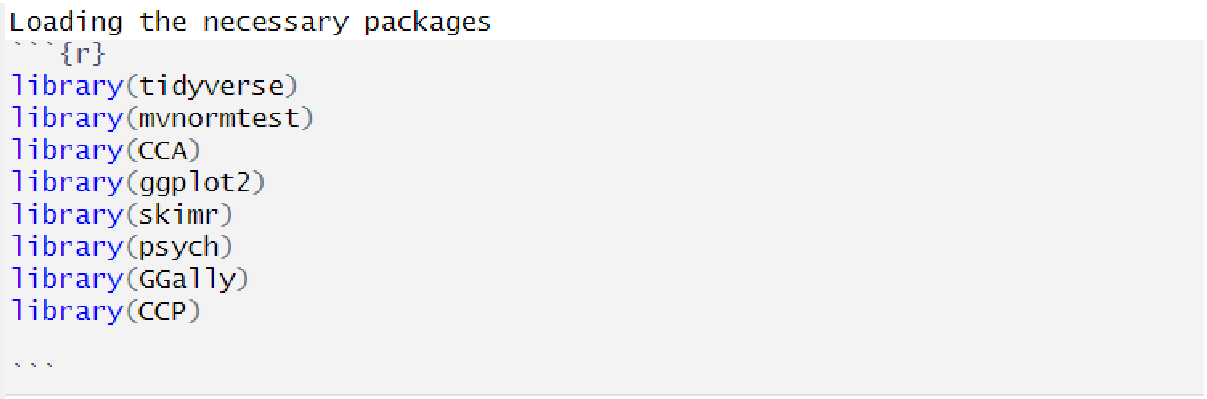


```
head(mmreg)
```

locus_of_control <dbl>	self_concept <dbl>	motivation <dbl>	read <dbl>	write <dbl>	math <dbl>	science <dbl>	female <dbl>
-0.84	-0.24	1.00	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
0.71	0.28	0.67	62.7	56.7	54.7	58.0	0
-0.64	0.03	1.00	41.6	46.3	38.4	36.3	1
1.11	0.90	0.33	62.7	64.5	61.4	58.0	1

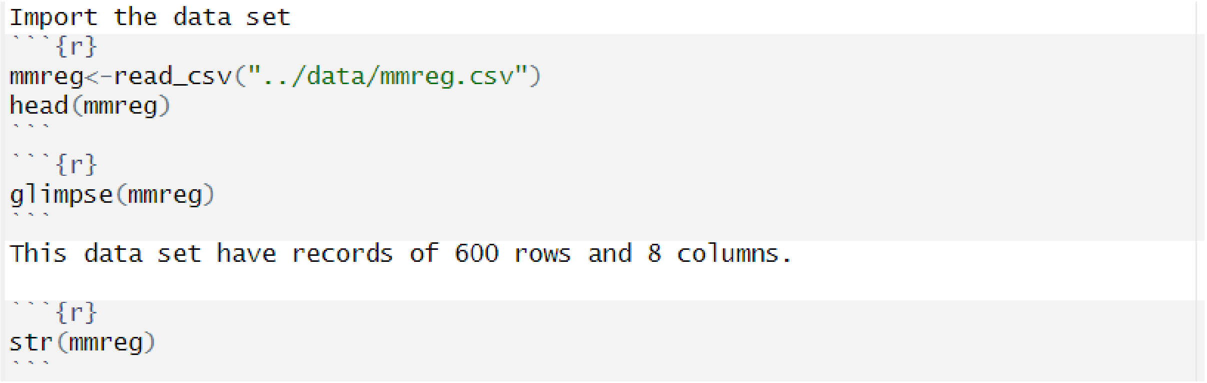
6 rows

- Codes

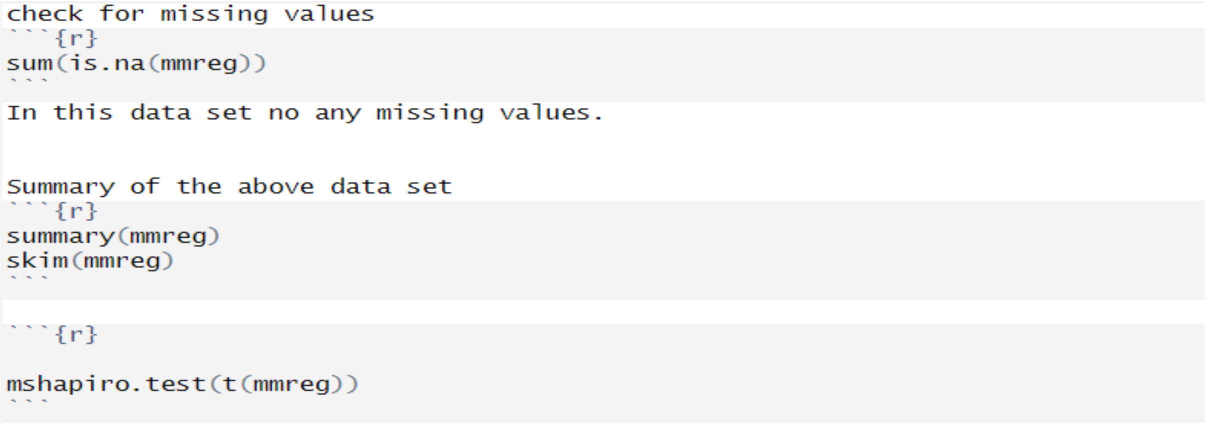
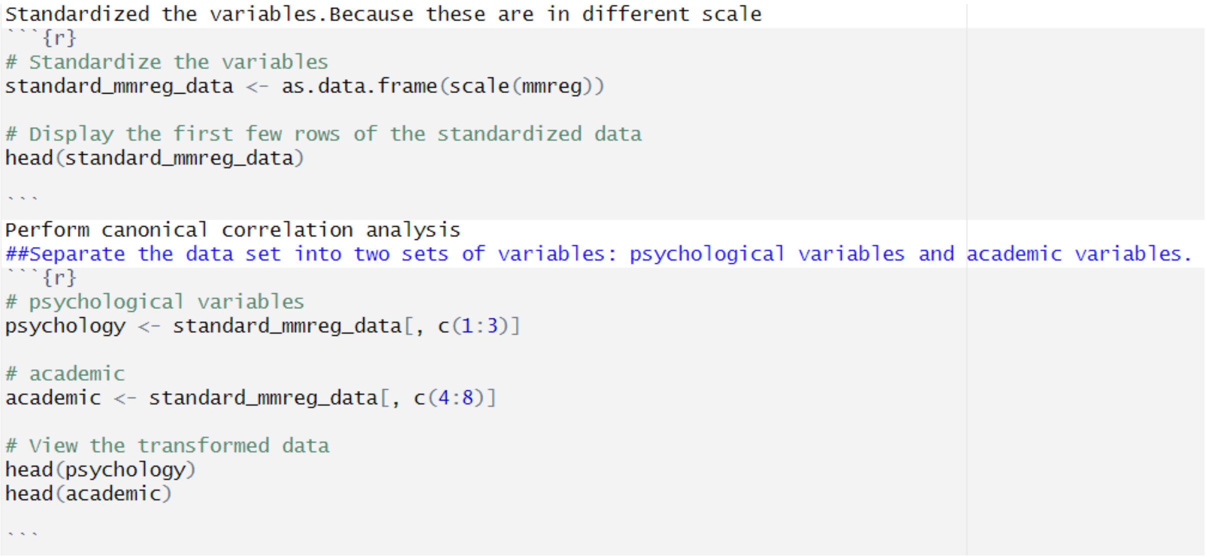


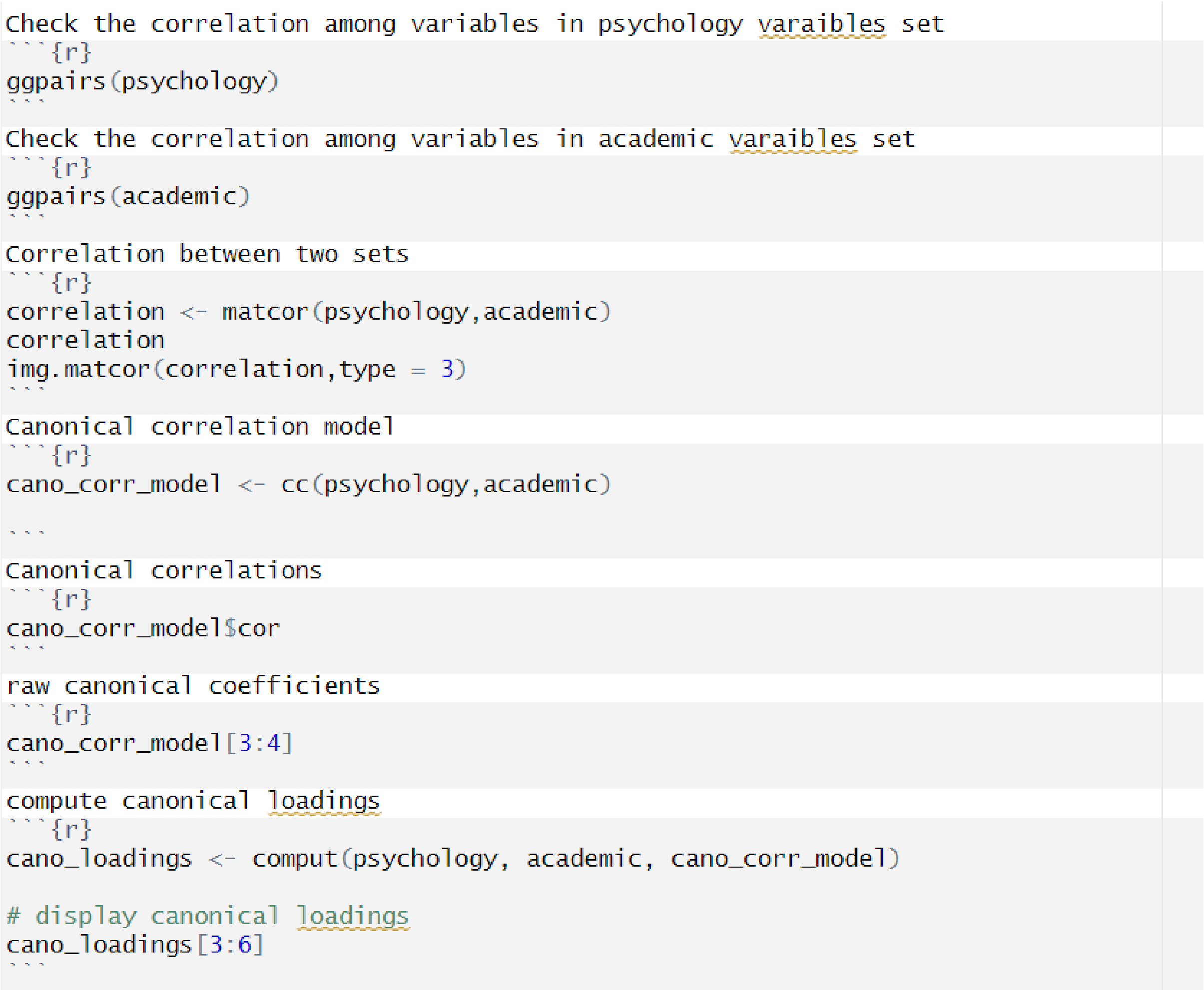
Loading the necessary packages

~~~
```{r}
library(tidyverse)
library(mvnormtest)
library(CCA)
library(ggplot2)
library(skimr)
library(psych)
library(GGally)
library(CCP)

```
~~~

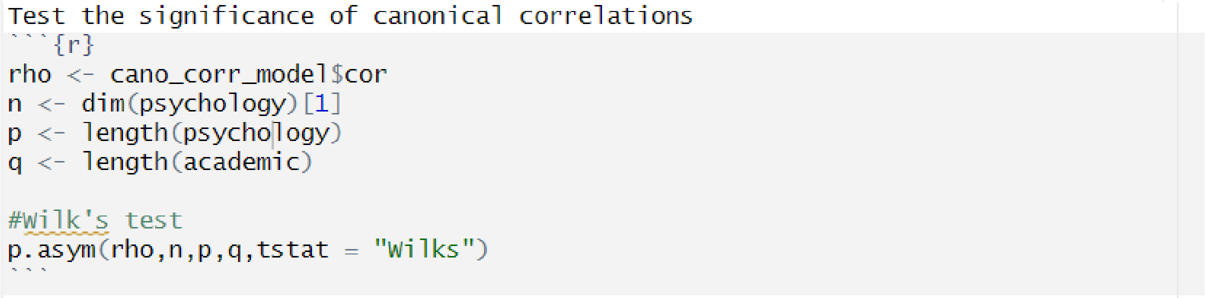


Import the data set

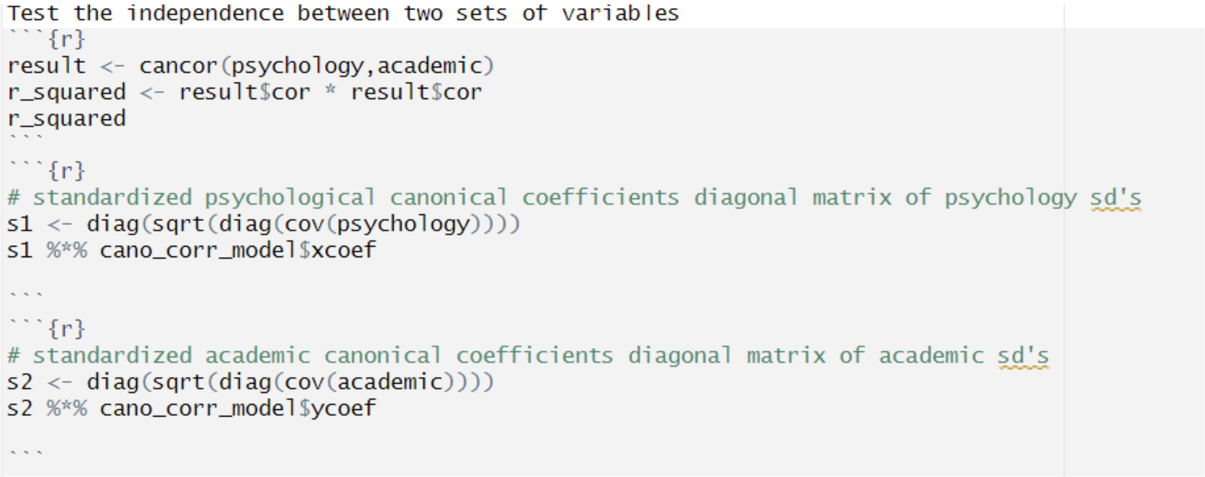
~~~
```{r}
mmreg<-read_csv("../data/mmreg.csv")
head(mmreg)
~~~
```{r}
glimpse(mmreg)
```
```

This data set have records of 600 rows and 8 columns.

```
```{r}
str(mmreg)
```
```



check for missing values

```
```{r}
sum(is.na(mmreg))
```
```

In this data set no any missing values.

Summary of the above data set

```
```{r}
summary(mmreg)
skim(mmreg)
```
```

```
```{r}

mshapiro.test(t(mmreg))
```
```



Standardized the variables.Because these are in different scale

```
```{r}
# Standardize the variables
standard_mmreg_data <- as.data.frame(scale(mmreg))

# Display the first few rows of the standardized data
head(standard_mmreg_data)

```
```

Perform canonical correlation analysis

##Separate the data set into two sets of variables: psychological variables and academic variables.

```
```{r}
# psychological variables
psychology <- standard_mmreg_data[, c(1:3)]

# academic
academic <- standard_mmreg_data[, c(4:8)]

# View the transformed data
head(psychology)
head(academic)

```
```
```



~~~
Check the correlation among variables in psychology varaibles set
```{r}
ggpairs(psychology)
```
Check the correlation among variables in academic varaibles set
```{r}
ggpairs(academic)
```
Correlation between two sets
```{r}
correlation <- matcor(psychology,academic)
correlation
img.matcor(correlation,type = 3)
```
Canonical correlation model
```{r}
cano_corr_model <- cc(psychology,academic)

```
Canonical correlations
```{r}
cano_corr_model$cor
```
raw canonical coefficients
```{r}
cano_corr_model[3:4]
```
compute canonical loadings
```{r}
cano_loadings <- comput(psychology, academic, cano_corr_model)

# display canonical loadings
cano_loadings[3:6]
```
~~~



~~~
Test the significance of canonical correlations
```{r}
rho <- cano_corr_model$cor
n <- dim(psychology)[1]
p <- length(psychology)
q <- length(academic)

#Wilk's test
p.asym(rho,n,p,q,tstat = "Wilks")
```
~~~



~~~
Test the independence between two sets of variables
```{r}
result <- cancor(psychology,academic)
r_squared <- result$cor * result$cor
r_squared
```
```{r}
# standardized psychological canonical coefficients diagonal matrix of psychology sd's
s1 <- diag(sqrt(diag(cov(psychology))))
s1 %*% cano_corr_model$xcoef

```
```{r}
# standardized academic canonical coefficients diagonal matrix of academic sd's
s2 <- diag(sqrt(diag(cov(academic))))
s2 %*% cano_corr_model$ycoef

```
~~~