MSDS 6372 PROJECT 2

Team Members:

Adesanya Olufemi

Anderwald Scott

Scott Marvin

Date: 11/06/2016

**Canonical Correlation (CCA) Analysis**

**Introduction**

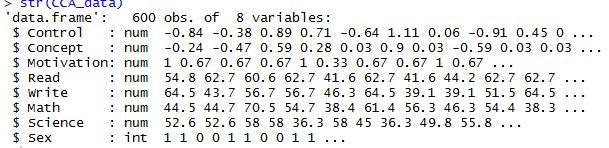
Canonical correlation analysis is used to identify and measure the associations among two sets of variables. It is appropriate in the same situations where multiple regression would be, but where there are multiple intercorrelated outcome variables. Canonical correlation analysis determines a set of canonical variates, orthogonal linear combinations of the variables within each set that best explain the variability both within and between sets.

For this section of our paper, we will be conducting a CCA analysis, we are interested in how the set of psychological variables relates to the academic variables and gender. In particular, we are interested in how many dimensions (canonical variables) are necessary to understand the association between the two sets of variables. Our dataset used for this analysis include 600 observations and 13 variables.

**Exploratory Data Analysis**

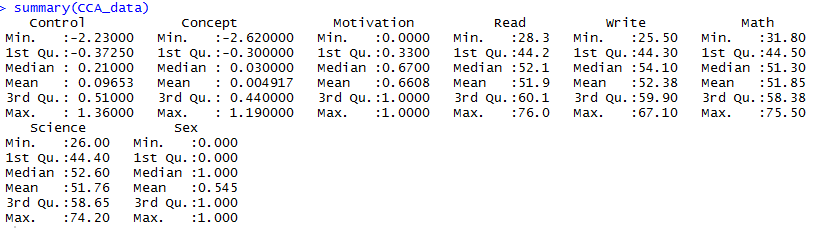
For this analysis, we investigated the association between psychological measures and academic achievement measures. Our data (mmreg.csv) include 600 observation on eight variables. The psychological variables are locus\_of\_control, self\_concept and motivation. The academic variables are read, write, math, and science while the variable female is a zero-one indicator variable with the one indicating a female student. The first step we took was to see the structure of our dataset using R. Fig 1.0 shows the structure of our dataset

Fig 1.0 Dataset Structure



Furthermore we did a summary statistics in our dataset for each variable, Fig 1.1 below shows the output

Fig 1.1 Summary Statistics



Next, we conducted our Canonical correlation analysis, this requires two sets of variables. We specify our psychological variables as the first set of variables and our academic variables plus gender as the second set. Fig 1.2 below shows the output:

Fig 1.2 Summary of both psychological and academic groups

|  |  |
| --- | --- |
| Psychological Variables | Academic variables |
|  |  |

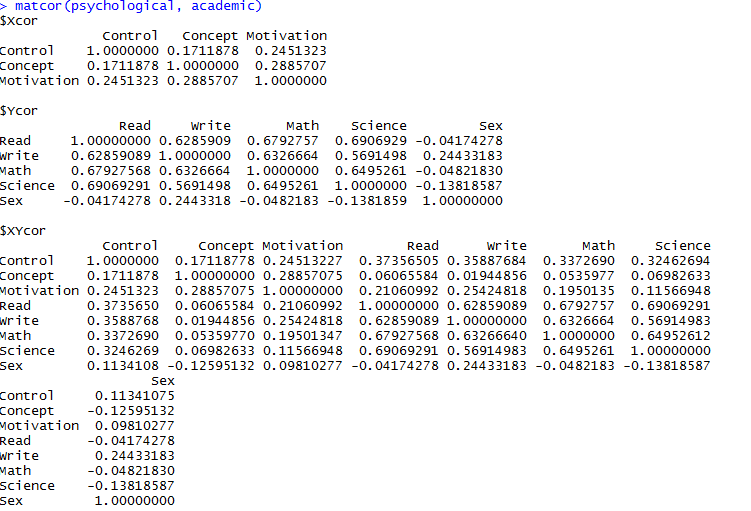
Next, we plotted a graph to show the correlation for each variable group, Fig 1.3 shows the output for the psychological and Academic variables group

Fig 1.3 Correlation Plots for Psychological and Academic variable groups

|  |  |
| --- | --- |
| **Psychological variables** | **Academic variables** |
|  |  |

Next, we look at the correlations within the two sets of variables using the CCA function in R. Fig 1.4 below shows the output of the correlation between the two sets of variables

Fig 1.4 correlation between within psychological and academic variables



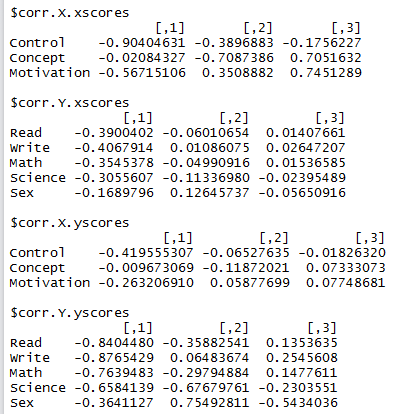
Next, we perform a CCA analysis using two canonical correlation analysis methods. Our first CCA method is CC1 while the second method is CC2. Fig 1.5 shows the output for our CC1 method below, from the below output for the raw canonical coefficients, for the variable read, a one unit increase in reading leads to a .0446 decrease in the first canonical variate of set 2 when all of the other variables are held constant.

Fig 1.5 CC1 displays the canonical coefficients between the variable groups and the raw canonical coefficients

|  |  |
| --- | --- |
| Canonical Coefficients | Raw canonical coefficients |
|  |  |

Next, we used another method of CCA we called CC2 for our analysis, we analyzed the correlation between variables and the canonical variates. The output below shows a correlation between observed variables and canonical variables, Fig 1.6 below shows our analysis output

Fig 1.6 CC2 analysis: correlations between variables and the canonical variates



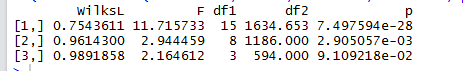
Next, we did a standardized coefficient analysis for easier comparison among the variables in the two groups. Fig 1.7 below shows the output for both psychological and academic group. In the output below, consider the variable “read”, a one standard deviation increase in reading leads to a 0.45 standard deviation decrease in the score on the first canonical variate for set 2 when the other variables in the model are held constant.

Fig 1.7 Standardized canonical coefficient

|  |  |
| --- | --- |
| Standardized psychological canonical coefficients | Standardized academic canonical coefficients |
|  |  |

Next, we did a test of dimensionality for our canonical correlation analysis, Fig 1.8 below shows our output

Fig 1.8 Test of Dimensionality



**Conclusion**

Based on our analysis using the canonical correlation analysis, below are our findings based on the two methods used:

Test of Dimensionality for the canonical correlation analysis indicates that two of the three canonical dimensions are statistically significant at the .05 level. Dimension 1 has a canonical correlation of 0.46 between the sets of variables, while dimension 2 the canonical correlation was much lower at 0.17. Fig 1.9 below shows the output

Fig 1.9 Test of Dimensionality

|  |
| --- |
| Tests of Canonical Dimensions  Canonical Mult.  Dimension Corr. F df1 df2 p  1 0.46 11.72 15 1634.7 0.0000  2 0.17 2.94 8 1186 0.0029  3 0.10 2.16 3 594 0.0911 |

Fig 1.10 below presents the standardized canonical coefficients for the first two dimensions across both sets of variables. For the psychological variables, the first canonical dimension is most strongly influenced by locus of control (.84) and for the second dimension self-concept (-.84) and motivation (.69). For the academic variables plus gender, the first dimension was comprised of reading (.45), writing (.35) and gender (.32). For the second dimension writing (.41), science (-.83) and gender (.54) were the dominating variables.

Fig 1.10 Standardized Coefficient

|  |
| --- |
| Standardized Canonical Coefficients  Dimension  1 2  Psychological Variables  locus of control -0.84 -0.42  self-concept 0.25 -0.84  motivation -0.43 0.69  Academic Variables plus Gender  reading -0.45 -0.05  writing -0.35 0.41  math -0.22 0.04  science -0.05 -0.83  gender (female=1) -0.32 0.54 |

**MANOVA Example**

**Two Mean Vectors with Three Groups:**

manova.data <- data.frame(group = as.factor(rep(1:3,c(4, 3, 5))), y1 = c(2, 3, 5, 2, 4, 5, 6,

7, 8, 10, 9, 7), y2 = c(3, 4, 4, 5, 8, 6, 7, 6, 7, 8, 5, 6))

y\_1<-with(manova.data, tapply(y1, group, mean))

**1 2 3**

**3.0 5.0 8.2**

y\_2<-with(manova.data, tapply(y2, group, mean))

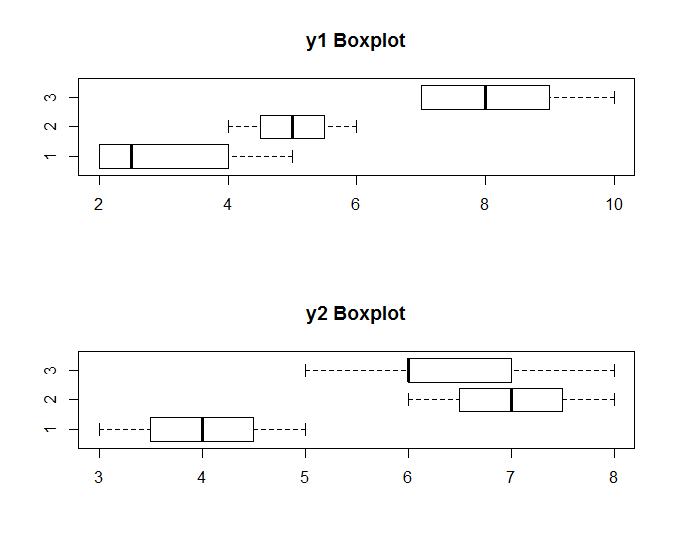
**1 2 3**

**4.0 7.0 6.4**

par(mfrow = c(2, 1))

boxplot(y1 ~ group, manova.data, main = "y1 Boxplot",horizontal = T)

boxplot(y2 ~ group, manova.data, main = "y2 Boxplot",horizontal = T)



m1 <- manova(cbind(y1, y2) ~ group, manova.data)

summary(m1)

**Df Pillai approx F num Df den Df Pr(>F)**

**group 2 1.3018 8.3899 4 18 0.0005283 \*\*\***

**Residuals 9**

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Conclusion:** Enough evidence to reject null hypothesis that no difference exists in any pair of the mean vectors.

install.packages("CCA")

#CCA ANALYSIS#

CCA\_data <- read.csv("http://www.ats.ucla.edu/stat/data/mmreg.csv") #Raw data import

colnames(CCA\_data) <-c("Control", "Concept", "Motivation", "Read", "Write", "Math", "Science", "Sex") #RENAME COLUMNS

summary(CCA\_data) #sUMMARY sTATISTICS

str(CCA\_data)

xtabs(~Sex, data =CCA\_data) #Sex variable data summary

psychological <- CCA\_data[, 1:3] #Grouping Psychological variables which include Control, Concept, MotivationVariables

academic <- CCA\_data[, 4:8] #Grouping the academic variables which include Read, Write, Math, Science, Sex Variables

require("GGally")

ggpairs(psychological) #Corellation Plot for psychological variables

require("GGally")

ggpairs(academic) #Corellation Plot for academic variables

# correlations using CCA

require ("CCA")

matcor(psychological, academic)

cc1 <- cc(psychological, academic)

# display the canonical correlations

cc1$cor

# raw canonical coefficients

cc1[3:4]

# compute canonical loadings

cc2 <- comput(psychological, academic, cc1)

# display canonical loadings

cc2[3:6]

# standardized psych canonical coefficients diagonal matrix of psych sd's

scoef <- diag(sqrt(diag(cov(psychological))))

scoef %\*% cc1$xcoef

# standardized acad canonical coefficients diagonal matrix of acad sd's

scoef2 <- diag(sqrt(diag(cov(academic))))

scoef2 %\*% cc1$ycoef

# tests of canonical dimensions

ev <- (1 - cc1$cor^2)

n <- dim(psych)[1]

p <- length(psych)

q <- length(acad)

k <- min(p, q)

m <- n - 3/2 - (p + q)/2

w <- rev(cumprod(rev(ev)))

# initialize

d1 <- d2 <- f <- vector("numeric", k)

for (i in 1:k) {

s <- sqrt((p^2 \* q^2 - 4)/(p^2 + q^2 - 5))

si <- 1/s

d1[i] <- p \* q

d2[i] <- m \* s - p \* q/2 + 1

r <- (1 - w[i]^si)/w[i]^si

f[i] <- r \* d2[i]/d1[i]

p <- p - 1

q <- q - 1

}

pv <- pf(f, d1, d2, lower.tail = FALSE)

(dmat <- cbind(WilksL = w, F = f, df1 = d1, df2 = d2, p = pv))

Introduction for Principal Component Analysis

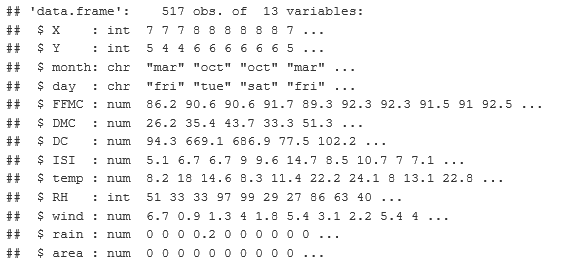
A major concern of any area that is heavily forested is the threat of a wildfire. This threat affects both wildlife and any structures in the area. Varies inputs can be used to help identify and mitigate this risk. This would help government agencies determine where firefighting resources can be placed to contain any major forest fire.

The principal component analysis was used for this section of the project. The purpose of the principal analysis component (pca)is to simplify the description of a set of interrelated varialbes. Along with the simplification of the variables is a reduction in noise, and helps identify how different varialbes work together. The data set was obtained from the country of Portugal with inputs from individual weathers indicators and the Canadian fire system

Exploratory Data Analysis

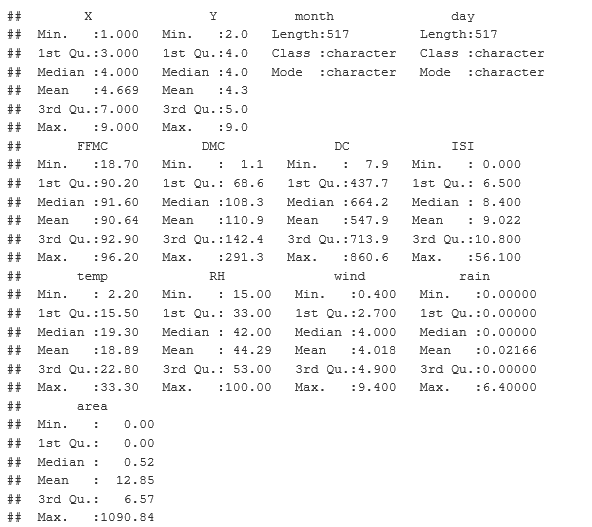
For this study, we investigated which input would be most beneficial for the detection of forest fires. The data set contains 517 observations and seven variables. The variables for the dataset include a coded fire danger system from Canada. This system takes input from climate and fire load to output a numeric system. Along the Canadian system individual weather inputs were used for the analysis. As with any analysis the first step should include a look at the data and the determine the structure. See figure below.

Data structure



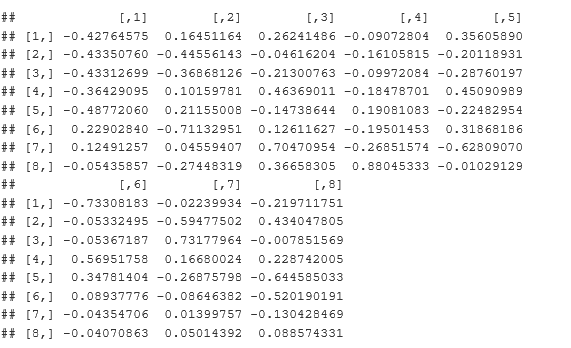
For this analysis, certain variables were removed this include X, Y, month, and year. X and Y were coordinates for data collection and was not provided to us. Further EDA included summary of the variables used in the analysis. Next a summary of the data was made to determine the range within the dataset. During the analysis outliers were noted and were included into the analysis. While normally outliers would normally be removed, we believed outliers were needed to give a full range of data. See below for the data summary.

Data summary



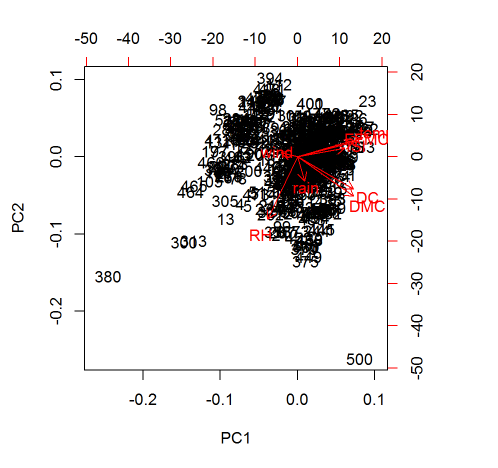
Next we began the process of PCA in R to reduce the dimensionality of the data. We include the following variables from both the Canadian fire system and the individual weather inputs. See below for the eigen values

Table of Eigen values



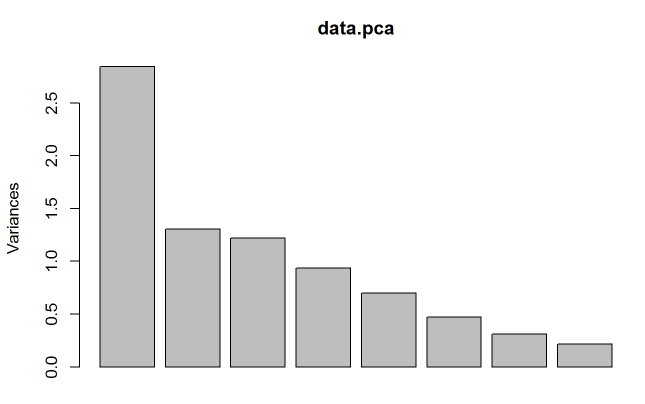
Next we created a biplot showing the principal components and the all the observations. See figure below.

Biplot

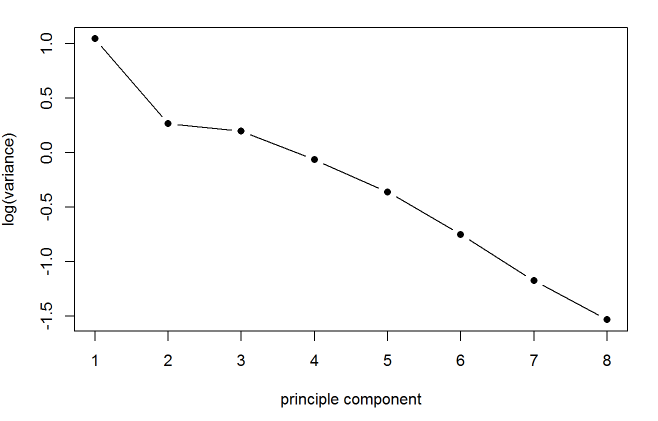


Next we calculated the correlation coefficients between variables and principal components. After this we plotted the scree graphs which indicate eigenvalues in non-increasing order of variance. Typically any value with an eigen value above 1.0 help determine the most important parameters. See figures below for the scree graphs.

Variance versus component plot



Log(variances) versus principal component plot



Conclusion

Based on your analysis using the principal Component Analysis of the fire data, we can conclude there is three variables which provide the greatest amount of variance in the dataset. Please see below for the correlation matrix. These three values have the highest amount of variance of the seven inputs.

Correlation matrxi table

