**Canonical Correlation Analysis**

**(CCA)**

**ST405 Multivariate Methods II**

**Assignment II**

S/18840

Dinuka Weerasooriya

# 1 Introduction

In multivariate statistics, Canonical Correlation Analysis (CCA) technique is used to identify and measure the associations between two sets of variables. Unlike simple correlation which deals with the relationship between two variables, CCA extends this concept to explore relationships between multiple interrelated quantitative variables, thus providing deeper understanding about the data set.

In this study, I'll analyze mentioned dataset using canonical correlation analysis. Let's explore how it applies to our selected data.

### Purpose of the study

The purpose of canonical correlation analysis is dimension reduction. I am going to try to determine how many numbers of canonical variate pairs should we consider and the values of canonical variate scores, and relationships between variables in the first set and those in the second set.

### Hypothesis

In this study, we need to test the hypothesis that at least the first canonical correlation is statistically significant. We will repeat this process until we find a non-significant canonical correlation. Additionally, we need to conduct a hypothesis test to determine the relationship between the first and second data sets.

### Importance of the study

In lots of studies, we have large number of individual variables, and handling big datasets can be tough. That's why finding a efficient way to reduce the data set’s dimensions is important. Canonical correlation analysis helps with issue, making it easier to make predictions and conclusions.

# Methodology

In this study, we applied Canonical Correlation Analysis (CCA) to a dataset from the World Health Organization (WHO) and the United Nations, which includes health and economic indicators for 193 countries from the years 2000 to 2015. The dataset consists of 22 variables, capturing various aspects of health and economic status.

Here is the description of each variables,

|  |  |
| --- | --- |
| Variable name | Description |
| "Country" | Country |
| "Year" | Year |
| "Status" | Developed or Developing status |
| "Life expectancy" | Life Expectancy in age |
| "Adult Mortality" | Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population) |
| "infant deaths" | Number of Infant Deaths per 1000 population |
| "Alcohol" | Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol) |
| "percentage expenditure" | Expenditure on health as a percentage of Gross Domestic Product per capita(%) |
| "Hepatitis B" | Hepatitis B (HepB) immunization coverage among 1-year-olds (%) |
| "Measles" | Measles - number of reported cases per 1000 population |
| "BMI" | Average Body Mass Index of entire population |
| "under-five deaths" | Number of under-five deaths per 1000 population |
| "Polio" | Polio (Pol3) immunization coverage among 1-year-olds (%) |
| "Total expenditure" | General government expenditure on health as a percentage of total government expenditure (%) |
| "Diphtheria" | Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%) |
| "HIV/AIDS" | Deaths per 1 000 live births HIV/AIDS (0-4 years) |
| "GDP" | Gross Domestic Product per capita (in USD) |
| "Population" | Population of the country |
| "thinness 1-19 years" | Prevalence of thinness among children and adolescents for Age 10 to 19 (% ) |
| "thinness 5-9 years" | Prevalence of thinness among children for Age 5 to 9(%) |
| "Income composition of resources" | Human Development Index in terms of income composition of resources (index ranging from 0 to 1) |
| "Schooling" | Number of years of Schooling(years) |

For the canonical correlation analysis purpose, I selected only two sets of variables. From here I called my first variable set as “Economic and social” variables and the second variable set as “Health related” variables.

Since there are large number of variables in the data set and to make it easier for the analysis I have chosen only 3 most important variables for each set. So I have included the variables of each set as follows

|  |  |
| --- | --- |
| **Economic and social** | **Health related** |
| GDP | Life expectancy |
| Population | Adult Mortality |
| percentage expenditure | Infant deaths |

Since different variables have different measurement units, I standardized the whole data set

In this study I’m going to apply statistical methods to perform the canonical correlation analysis on above mentioned two variable sets using R software. For the convenience, the variables in the first set are referred as “eco\_social” variables and the variables in the second set are called “health” variables in the R

# Results and Discussions

### Pairwise Scatter plots

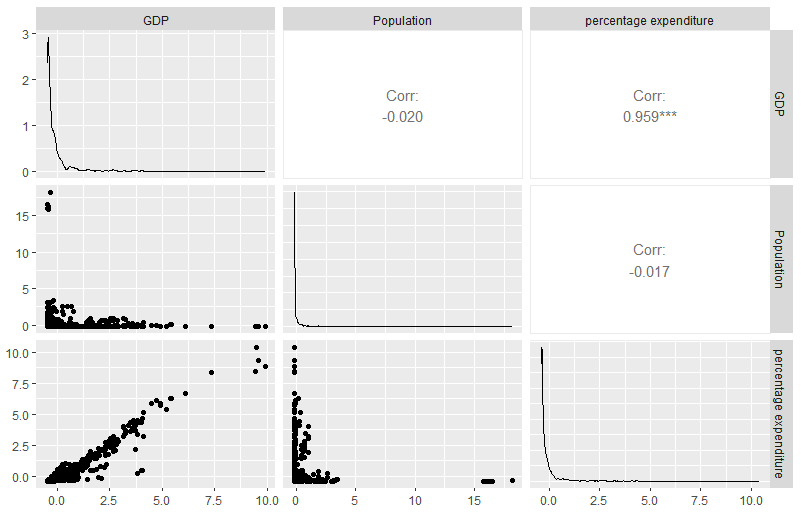


Figure 1 Pairwise scatter plot for the eco\_social variables

### 

Figure 2 Pairwise Scatter plot for the health variables

### 3.2Correlation Matrix

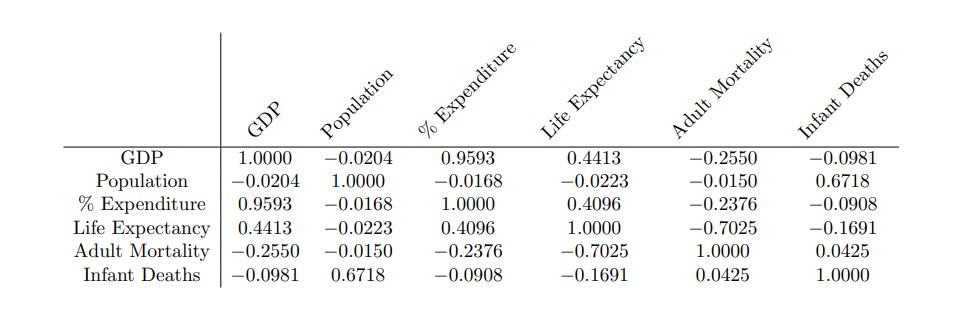
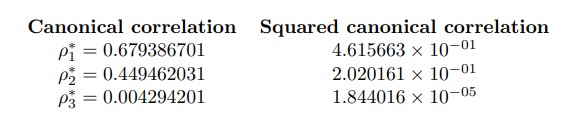


Figure 3 correlation matrix of six variables chosen to CCA

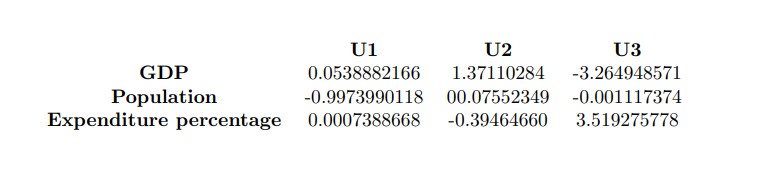
### 3.3 Canonical correlations and squared canonical correlations



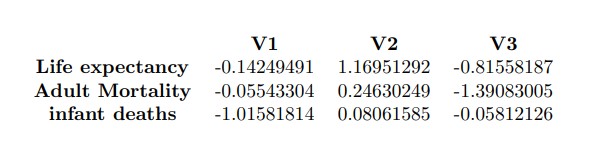
### 3.4 Estimated canonical coefficients

### 

#### 3.4.1. Estimated canonical coefficients(aij) for the Economic and social variables

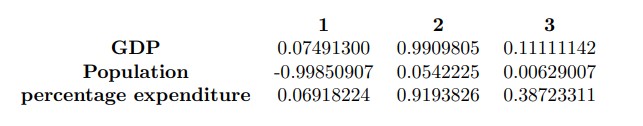


#### 3.4.2. Estimated canonical coefficients(bij) for the health variables

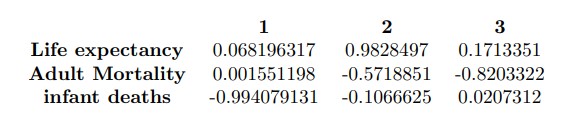


### 3.5. The correlations between each variable and the canonical variables

#### 3.5.1. The correlations between the square Economic and social variables and the canonical variables of itself



#### 3.5.2. The correlations between the health variables and the canonical variables of itself



### 3.6. Testing the relationship between the canonical variate pairs

In here we wish to test for the independence between the two sets of variables.

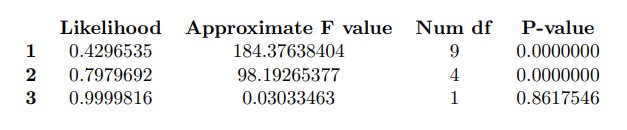
Hypothesis;

H01 : First set of variables is independent from the second set of variables

vs

H11 : First set of variables is dependent from the second set of variables

For this purpose, we can use likelihood ratio test. It is carried out using Wilk’s lambda.



1. First Canonical Correlation:

* Wilk's lambda= 0.4297; F= 184.376; p-value = 0.0000
* Therefore, we reject the null hypothesis that there is no relationship between the two sets of variables.
* We can conclude that the two sets of variables are dependent.
* The above hypothesis is also equivalent to testing all the canonical variate pairs are uncorrelated. (Hypothesis; H02 : ρ1\* = ρ2\* = ρ3\*= 0 vs. H12 : At least ρ1\* ≠ 0)
* Therefore, by considering this hypothesis, we can also say that at least the first canonical correlation is significant.
* Next we have to check whether the second and third canonical variate pairs are correlated or not. (Hypothesis; H03 : ρ2\* = ρ3\*= 0 vs. H13 : At least ρ2\* ≠ 0).

2. Second Canonical Correlation:

* + Wilk's lambda = 0.7980; F= 98.193; p-value = 0.0000
  + We reject H03 and this implies that at least the second canonical correlation is significant.
* Finally, we have to check whether the third canonical variate pair is correlated or not. (Hypothesis; H04 : ρ3\*= 0 vs. H14 : ρ3\* ≠ 0).

3. Third Canonical Correlation:

* + Wilk's lambda= 0.99998; F = 0.0303; p-value = 0.8618
  + We fail to reject H04 at the 5% significance level, implying that the third canonical correlation is not significant.
* We can test above each and every hypothesis using Pillai’s test, Lawley-Hotelling test and Roy’s largest root test. Those are also provided similar results.

# Conclusions and recommendations

•By considering squared values of the canonical variate pairs we can conclude that 46.15% of the total variation in U1 is explained by the variation in V1, and 20.20% of the variation in U2 is explained by V2. But only 0.000018% of the variation in U3 is explained by V3.

•The first one is a moderate high canonical correlation and implies that first canonical correlation is very important rather than the others.

•Using 3.4.1, the first canonical variable for square “Economic and social” is,

**U1 = - 0.053882GDP – 0.99736Population +0.0007Expenditure**

•Using 3.4.2, the first canonical variable for “Health” is,

**V1 = -0.14249491Life\_expectancy– 0.05543Adult\_mortality – 1.0158infant\_deaths**

• The magnitudes of the above coefficients give the contributions of the individual variables to the corresponding canonical variable.

• By considering the correlations between **“Economic and social”** variables and its canonical variables, for the first canonical variable, **“Population”** variable has a large negative correlation (-0.9985). Other two variables have moderately low positve correlations.Therefore we can conclude that there is a very strong inverse relationship between Population and first canonical variable

• Second and third canonical variables for **“Economic and social”**, all correlations have a positive values. **GPD** and **Percentage expenditure** is Strongly associated with second canonical variable while third canonical variable does not show any strong associations.

• By considering the correlations between **“Health”** variables and its canonical variables, for the first canonical variable, **Infant Deaths** has a very high negative correlation (-0.9941) with the first canonical variable, indicating a very strong inverse relationship between Infant Deaths and this canonical variable. While other two variables have a very weak associations

• When you consider Second and third canonical variables for **“Health”**, **Adult Mortality** is most strongly (and negatively) associated with the third canonical variable**.Infant Deaths** is most strongly (and negatively) associated with the first canonical variable

•As a summary, we can say that both variable sets are dependent. Not All three canonical correlations are significant but first two are Significant. Canonical correlation analysis technique can be easily applicable to this analysis.

# References

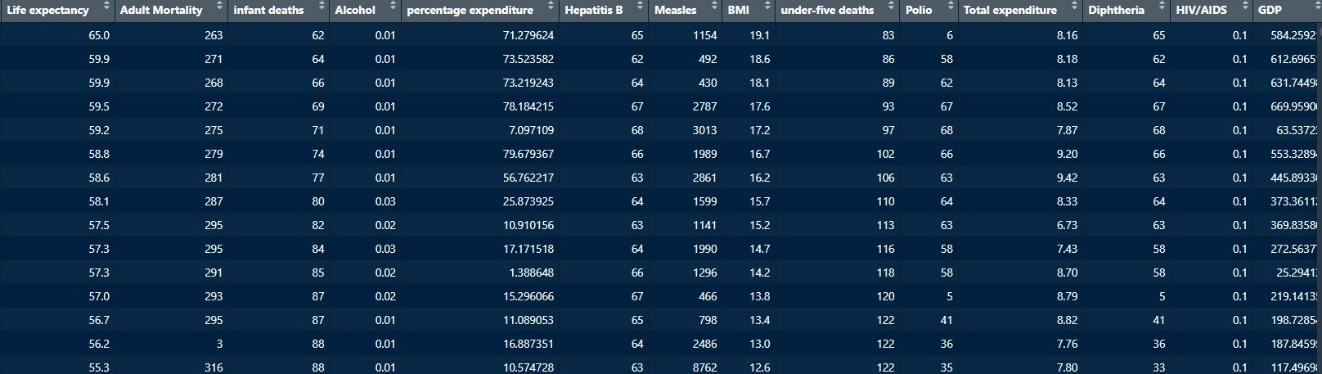
Kumar, A. (n.d.). Life expectancy (WHO). Kaggle. Retrieved May 16, 2024, from <https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who/data>

González, I., Déjean, S., Martin, P. G. P., & Baccini, A. (2023). CCA: Canonical correlation analysis. R package version 1.2. Retrieved from <https://cran.r-project.org/web/packages/CCA/CCA.pdf>

Devy. (2022). Canonical Correlation Analysis with R. RPubs. Retrieved May 20, 2024, from <https://rpubs.com/Devy/902673>

# Appendices

#head of the data



#Loading Libraries that are needed

> library(dplyr)

> library(tidyr)

> library(Amelia)

> library(tidyverse)

> require(GGally)

> require(CCA)

> require(CCP)

#importing the data Set

> life\_exp\_data<-read\_csv("../data/Life Expectancy Data.csv")

**Rows:** 2938 **Columns:** 22── **Column specification**

#Cleaning the dataset

> life\_exp\_data <- life\_exp\_data %>% select\_if(~ mean(is.na(.)) < 0.5)

> life\_exp\_data<-life\_exp\_data%>%drop\_na()

# Standardize numeric columns

> life\_exp\_data <- life\_exp\_data%>% mutate\_if(is.numeric, scale)

# Select the variables for X and Y

> eco\_social<- as.matrix(life\_exp\_data[, c("GDP", "Population", "percentage expenditure")])

>

> health<- as.matrix(life\_exp\_data[, c("Life expectancy", "Adult Mortality", "infant deaths")])

#Pairwise scatter plots.

> ggpairs(eco\_social)

> ggpairs(health)

#Find the correlations within and between the two sets of variables.

> matcor(eco\_social,health)

#Apply Canonical Correlation Analysis to the two groups.

> cc1 <- cc(eco\_social,health)

#The canonical correlations.

> cc1$cor

#The squared canonical correlations

> (cc1$cor)^2

#Raw canonical coefficients for the eco\_social variables.

> cc1[3]

#Row canonical coefficients for the health variables.

> cc1[4]

#Compute canonical loadings.

> cc2 <- comput(eco\_social,health, cc1)

#Correlations between each variable and the corresponding canonical variate.

> cc2[3:6]

#Tests of canonical dimensions.

> rho <- cc1$cor

#Define number of observations, number of variables in first set, and number of variables in the second set.

> n <- dim(X)[1]

> p <- length(eco\_social)

> q <- length(health)

#Calculate p-values using the F-approximations of different test statistics:

> 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")

#Standardized square feet sizes canonical coefficients diagonal matrix of eco\_social standard deviations.

> s1 <- diag(sqrt(diag(cov(eco\_social))))

> s1 %\*% cc1$xcoef

#Standardized facilities canonical coefficients diagonal matrix of facilities standard deviations.

> s2 <- diag(sqrt(diag(cov(health))))

> s2 %\*% cc1$ycoef