

Canonical Correlation Analysis Report

ST405 – Assignment 02

S/18/805

1. Introduction

Canonical Correlation Analysis is a statistical technique used to identify and quantify the relationship between two multivariate sets of variables, all measured in same on the same individual. There are several variables in each of two sets. Canonical Correlation Analysis look for linear combination of variables which is known as canonical variables or canonical variates, within each set so that the correlation between them is maximized. The two datasets' covariation patterns are represented by these canonical variates while relationships' strength is measured by the canonical correlations.

Canonical Correlation Analysis enables a deeper understanding of the underlying patterns and linkages in multidimensional data.

- **Main objective**

Find the relationship and patterns of linkage between two groups of multivariate variables.

2. Methodology

Dataset used in this study is **Life Expectancy (WHO)** which is a result of statistical analysis on factors influencing life Expectancy. It contain data from year 2000-2015 for 193 countries. There are 22 columns (variables) and 2938 rows in the dataset. Variables can be divided into several broad categories such as immunization factors, mortality factors, economical factors and social factors.

- Variables

- 1 Country
- 2 Year
- 3 Status – Country is developed or developing
- 4 Lifeexpectancy – Life expectancy in age
- 5 AdultMortality – Adult mortality rates of both sexes (Probability of dying between age 16 and 60 per 1000 population)
- 6 infantdeaths – Number of infant deaths per 1000 population
- 7 Alcohol – Alcohol consumption in liters of pure alcohol (recorded per capita (15+))

- 8 percentageexpenditure – Expenditure on health as a percentage of gross domestic product per capita (%)
- 9 HepatitisB – Hepatitis B immunization coverage among 1 year olds (%)
- 10 Measles – number of reported cases per 1000 population
- 11 BMI – Average body mass index of entire population
- 12 under-fivedeaths – Number of under-five deaths per 1000 population
- 13 Polio – Polio immunization coverage among 1 year olds (%)
- 14 Totalexpenditure – General government expenditure on health as a percentage of total government expenditure
- 15 Diphtheria – Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1 year olds (%)
- 16 HIV/AIDS – Deaths per 1000 live births HIV/AIDS (0-4 years)
- 17 GDP – Gross Domestic Product per capita (in USD)
- 18 Population – Population of the country
- 19 thinness1-19years – Prevalence of thinness among children and adolescents for age 10 to 19 (%)
- 20 thinness5-9years - Prevalence of thinness among children for age 5 to 9 (%)
- 21 Incomecompositionofresources – Human development index in terms of income composition of resources (ranging from 0 to 1)
- 22 Schooling – Number of years of Schooling

Dataset contain missing values and they were removed. Also, Country, Year and Status variables are removed. After all pre-processing resultant dataset have 19 columns and 1649 rows.

Since variables are measured on different scales dataset was standardized.

Two datasets to apply canonical correlation analysis as follows:

Set 1 – Mortality factors

Lifeexpectancy, AdultMortality, infantdeaths, under-fivedeaths, HIV/AIDS

Set 2 – Health related & social factors

Alcohol, percentageexpenditure, HepatitisB, Measles, BMI, Polio, Totalexpenditure, Diphtheria, GDP, Population, thinness1-19years, thinness5-9years, Incomecompositionofresources, Schooling

Set 1 has 5 variables while the set 2 has 14 variables. Hence, there are five canonical variate pairs.

➤ Purpose of the study

Find the linear combinations that maximize the correlations between the members of each canonical variate pair.

3. Results and Discussion

3.1 Estimates of Canonical Correlations

There are five canonical correlations corresponding to each of the five canonical variate pairs.

```
```{r}
mortal_var <- std_data[,c("Lifeexpectancy", "AdultMortality", "infantdeaths", "under-fivedeaths", "HIV/AIDS")]
health_var <- std_data[,c("Alcohol", "percentageexpenditure", "HepatitisB", "Measles", "BMI", "Polio",
 "Totalexpenditure", "Diphtheria", "GDP", "Population", "thinness1-19years",
 "thinness5-9years", "Incomecompositionofresources", "Schooling")]
```
```

```
```{r}
ccv <- cc(mortal_var, health_var)
```
```

```
```{r}
cor <- ccv$cor
cor
```
```

```
[1] 0.85947867 0.78029810 0.34455061 0.16283560 0.08924259
```

Both first and second canonical correlations are high which are 0.859 and 0.780 respectively. Third canonical correlation is 0.344 and fourth canonical correlation is 0.163. They are considered as low while the fifth canonical correlation which is 0.089 is very low.

By considering all the correlation values we can say that there is a relationship between two sets of variables which we categorized as mortality factors and health related & social factors.

To test the statistical significance of these canonical correlations Wilks' lambda test were performed.

Test of H0: The canonical correlations in the current row and all that follow are zero

```

      CanR LR test stat approx F numDF  denDF  Pr(> F)
1 0.85948      0.08698   74.337    70 7764.4 < 2.2e-16 ***
2 0.78030      0.33289   39.911    52 6318.9 < 2.2e-16 ***
3 0.34455      0.85108    7.514    36 4822.7 < 2.2e-16 ***
4 0.16284      0.96573    2.611    22 3266.0 5.906e-05 ***
5 0.08924      0.99204    1.312    10 1634.0  0.2183
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Null hypothesis is all 5 canonical variate pairs are uncorrelated vs alternative hypothesis is at least one canonical variate pair is significant.

Considering first result $\Lambda = 0.085948$; $p < 2.2e-16$ Wilk's lambda is significant. So we reject null hypothesis. Since the canonical correlations are ordered from largest to smallest we can conclude that first canonical variate pair is significantly correlated.

Likewise, all canonical variate pairs are significantly correlated except the fifth one.

Since only first four canonical variate pairs are significantly correlated and depend on one another we summarize the results for those pairs.

3.2 Significant canonical correlations

| canonical_correlation
<dbl> | Squared_correlation
<dbl> |
|--------------------------------|------------------------------|
| 0.8594787 | 0.73870359 |
| 0.7802981 | 0.60886513 |
| 0.3445506 | 0.11871513 |
| 0.1628356 | 0.02651543 |

From squared canonical correlations corresponding to significantly correlated canonical variate pairs we can list down following findings.

73.87% of the variation in first canonical variable of mortality factors (set 1) is explained in first canonical variable of health related & social factors (set 2).

60.88% of the variation in second canonical variable of mortality factors (set 1) is explained in second canonical variable of health related & social factors (set 2).

Percentage of the variation in first canonical variable of mortality factors (set 1) that is explained in first canonical variable of health related & social factors (set 2) is 11.87%.

2.65% of the variation in fourth canonical variable of mortality factors (set 1) is explained in fourth canonical variable of health related & social factors (set 2).

Looking at the percentage values explains others' variation we can say that only the first and second variate pairs are important.

3.3 Canonical Coefficients

The magnitudes of these coefficients indicates the extent which to which each individual variables contributes to the corresponding canonical variable.

 *Estimated canonical coefficients for the mortality factors*

| | Mortality_1
<dbl> | Mortality_2
<dbl> | Mortality_3
<dbl> | Mortality_4
<dbl> |
|------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| Lifeexpectancy | -0.93373432 | 0.82582605 | -0.6395168 | 0.3477835 |
| AdultMortality | -0.03636944 | 0.07035978 | -0.2523617 | 0.4037186 |
| infantdeaths | 2.41577658 | 2.73991824 | 10.5933334 | 7.2005229 |
| under-fivedeaths | -1.89939616 | -1.90316481 | -10.8724106 | -7.3203106 |
| HIV/AIDS | -0.26432364 | 0.26565154 | -0.6137183 | 0.8160713 |


These are the estimated canonical coefficients for the mortality factors.

First canonical variable for mortality can be written as:

$$\text{Mortality_1} = -0.9337(\text{Lifeexpectancy}) - 0.0364(\text{AdultMortality}) + 2.4158(\text{infantdeaths}) - 1.8994(\text{under-fivedeaths}) - 0.2643(\text{HIV/AIDS})$$

Other canonical variables for mortality set can be written in the same way.

“Infantdeath” variable gives the highest contribution to the corresponding first and second canonical variables while “under-fivedeaths” variable gives the highest contribution to the third and fourth canonical variables.

 *Estimated canonical coefficients for the health related & social factors*

| | H&S_1
<dbl> | H&S_2
<dbl> | H&S_3
<dbl> | H&S_4
<dbl> |
|------------------------------|----------------|----------------|----------------|----------------|
| Alcohol | 0.102788715 | -0.06556197 | -0.97098406 | 0.15281508 |
| percentageexpenditure | -0.128206645 | 0.05042221 | -0.04360421 | -0.07251997 |
| HepatitisB | -0.021409526 | -0.09051110 | 0.13646110 | -0.12685207 |
| Measles | 0.204418942 | 0.38600365 | 0.24516246 | 0.19956548 |
| BMI | -0.090139161 | 0.14360063 | 0.11048404 | -0.24687773 |
| Polio | -0.048504635 | 0.01918276 | 0.20246672 | 0.20043511 |
| Totalexpenditure | -0.020493824 | 0.01002160 | -0.21017313 | 0.46455257 |
| Diphtheria | -0.048850622 | 0.05902917 | 0.30622831 | 0.52459794 |
| GDP | 0.006979744 | 0.06101028 | -0.06517983 | -0.06186236 |
| Population | 0.301802993 | 0.53725366 | -0.01558526 | -0.24738685 |
| thinness1-19years | 0.063807738 | 0.11854053 | -0.27500419 | -0.72604926 |
| thinness5-9years | 0.170154629 | 0.14595933 | -0.03902449 | 1.09915834 |
| Incomecompositionofresources | -0.230577787 | 0.37026310 | 0.17826753 | -0.52492267 |
| Schooling | -0.400955811 | 0.26200493 | 0.04089787 | 0.47924136 |

These are the estimated canonical coefficients for the health related and social factors.

First canonical variable for health related and social factors can be written as:

$$\begin{aligned} \text{H\&S_1} = & 0.1028(\text{Alcohol}) - 0.1282(\text{percentageexpenditure}) - 0.0214(\text{HepatitisB}) \\ & + 0.2044(\text{Measles}) - 0.0901(\text{BMI}) - 0.0485(\text{Polio}) - 0.0205(\text{Totalexpenditure}) - \\ & 0.0489(\text{Diphtheria}) + 0.0069(\text{GDP}) + 0.3018(\text{Population}) + 0.0638(\text{thinness1-19years}) + \\ & 0.1702(\text{thinness5-9years}) - 0.2306(\text{Incomecompositionofresources}) - 0.4009(\text{Schooling}) \end{aligned}$$

Other canonical variables for health related and social set can be written in the same way.

To the first health related & social canonical variables “Schooling” variable gives the highest contribution while population gives the highest contribution to the second canonical variable. “Alcohol” and “thinness5-9years” variables contribute highly to the third and fourth canonical variables.

3.4 Correlation between variables and canonical variate

✚ *Correlations between each mortality variable and the corresponding canonical variate*

| | [,1] | [,2] | [,3] | [,4] |
|------------------|------------|------------|------------|------------|
| Lifeexpectancy | -0.7948985 | 0.5217324 | 0.2005715 | -0.2291233 |
| AdultMortality | 0.4619331 | -0.3620846 | -0.3476792 | 0.4725688 |
| infantdeaths | 0.6765456 | 0.7080524 | -0.1527539 | -0.1325059 |
| under-fivedeaths | 0.6810864 | 0.6789185 | -0.2160867 | -0.1686714 |
| HIV/AIDS | 0.2502763 | -0.2006230 | -0.5040058 | 0.7453828 |

“Lifeexpectancy”, “Infantdeaths”, “under-fivedeaths” correlations are relatively large when consider the first canonical variable. Therefore, we can say that the first canonical variable reflects mortality to a considerable extent.

Same as above explanation, some of variables have high correlation with other canonical variables while some have weak correlations.

✚ *Correlations between each health related & social variable and the corresponding canonical variate*

| | [,1] | [,2] | [,3] | [,4] |
|------------------------------|------------|-------------|-------------|--------------|
| Alcohol | -0.4270826 | 0.21437251 | -0.22279171 | 0.022201652 |
| percentageexpenditure | -0.3929338 | 0.22301045 | -0.10316655 | -0.009988053 |
| HepatitisB | -0.2603929 | -0.04428667 | 0.11937614 | 0.043312146 |
| Measles | 0.3692793 | 0.41650222 | 0.06353520 | 0.018840569 |
| BMI | -0.5439964 | 0.18539419 | 0.02077575 | -0.040980633 |
| Polio | -0.3238171 | 0.12334826 | 0.10590446 | 0.068106562 |
| Totalexpenditure | -0.2494960 | 0.02459290 | -0.08814962 | 0.070728685 |
| Diphtheria | -0.3327662 | 0.13327343 | 0.12760462 | 0.086195446 |
| GDP | -0.4206332 | 0.23998189 | -0.09981947 | -0.009537140 |
| Population | 0.4004456 | 0.56012532 | -0.01016182 | -0.021233975 |
| thinness1-19years | 0.6086757 | 0.07206664 | -0.02604574 | 0.025986746 |
| thinness5-9years | 0.6061489 | 0.07678177 | -0.02522949 | 0.047979031 |
| Incomecompositionofresources | -0.6357485 | 0.41097614 | -0.01430333 | -0.016792347 |
| Schooling | -0.6966881 | 0.35776448 | -0.04264515 | 0.021039407 |

Several health related & social variables have moderate correlation with the corresponding first canonical variable. Therefore, we can say it measure health related & social involvement to some extent.

But, there are no considerable correlation among individual variables and other canonical variables. Second, third and fourth canonical variables yields little information about relevant data.

3.5 Reinforcing the Results

✚ *Correlations between each mortality variable and the canonical variates for Health related & social factors*

| | [,1] | [,2] | [,3] | [,4] |
|------------------|------------|------------|-------------|-------------|
| Lifeexpectancy | -0.6831983 | 0.4071068 | 0.06910703 | -0.03730943 |
| AdultMortality | 0.3970216 | -0.2825340 | -0.11979309 | 0.07695102 |
| infantdeaths | 0.5814766 | 0.5524919 | -0.05263146 | -0.02157668 |
| under-fivedeaths | 0.5853793 | 0.5297588 | -0.07445281 | -0.02746570 |
| HIV/AIDS | 0.2151071 | -0.1565458 | -0.17365551 | 0.12137485 |

Only the “Lifeexpectancy” has relatively high correlation while “Infantdeaths” and “under-fivedeaths” show moderate correlation with first canonical variate for health related & social factors. Because first canonical correlation is high.

Other canonical variables of health related & social do not have considerable correlation with mortality variables.

✚ *Correlations between each health related & social variable and the mortality canonical variates*

| | [,1] | [,2] | [,3] | [,4] |
|------------------------------|------------|-------------|-------------|-------------|
| Alcohol | -0.4969089 | 0.27473155 | -0.64661532 | 0.13634397 |
| percentageexpenditure | -0.4571769 | 0.28580160 | -0.29942350 | -0.06133827 |
| HepatitisB | -0.3029661 | -0.05675609 | 0.34646909 | 0.26598696 |
| Measles | 0.4296550 | 0.53377321 | 0.18440021 | 0.11570301 |
| BMI | -0.6329376 | 0.23759405 | 0.06029812 | -0.25166876 |
| Polio | -0.3767599 | 0.15807838 | 0.30736983 | 0.41825352 |
| Totalexpenditure | -0.2902875 | 0.03151731 | -0.25583938 | 0.43435640 |
| Diphtheria | -0.3871721 | 0.17079809 | 0.37035087 | 0.52934031 |
| GDP | -0.4894050 | 0.30755156 | -0.28970915 | -0.05856914 |
| Population | 0.4659168 | 0.71783504 | -0.02949298 | -0.13040131 |
| thinness1-19years | 0.7081918 | 0.09235783 | -0.07559337 | 0.15958885 |
| thinness5-9years | 0.7052518 | 0.09840056 | -0.07322433 | 0.29464706 |
| Incomecompositionofresources | -0.7396909 | 0.52669120 | -0.04151299 | -0.10312455 |
| Schooling | -0.8105939 | 0.45849718 | -0.12377034 | 0.12920643 |

With the first mortality canonical variate “Scooling” has a strong negative correlation. “Incomecompositionofresources” has relatively strong negative correlation while “thinness1-19years” and “thinness5-9years” has positive relatively strong correlation. Only the “Population” has relatively strong correlation with second canonical variate. None of others have significant correlation with second canonical variate.

4. Conclusion

We only need five canonical variate pairs to explain the dataset. We don’t need to consider each and every pairwise correlations between variables. Wilk’s lambda test results proved this statement.

73.87% of the variation in first canonical variable of mortality factors is explained in first canonical variable of health related & social factors while 60.88% of the variation in second canonical variable of mortality factors is explained in second canonical variable of health related & social factors.

Variables “infantdeaths”, “under-fivedeaths” from mortality set contribute more to the model.

From the five canonical correlations only the first four were significant. That is first canonical variate pairs are significantly correlated and depend on one another.

First four canonical correlations are 0.8595, 0.7803, 0.3445, and 0.1628. Cence, we can conclude that two sets of variables have strong relationship.

5. References

[How do you interpret canonical correlation](#)

[Package CCA](#)

Dataset: [Life expectancy \(WHO\)](#)

6. Appendices

| | Country | Year | Status | Lifexpectancy | AdultMortality | infantdeaths |
|---|-------------|------|------------|---------------|----------------|--------------|
| 1 | Afghanistan | 2015 | Developing | 65.0 | 263 | 62 |
| 2 | Afghanistan | 2014 | Developing | 59.9 | 271 | 64 |
| 3 | Afghanistan | 2013 | Developing | 59.9 | 268 | 66 |
| 4 | Afghanistan | 2012 | Developing | 59.5 | 272 | 69 |
| 5 | Afghanistan | 2011 | Developing | 59.2 | 275 | 71 |

```

library(tidyverse)

library(skimr)
library(CCA)
library(CCP)
library(candisc)

data <- read_csv("../Data/led.csv")
skim(data)
clean_data <- na.omit(data)
clean_data <- clean_data[,-c(1,2,3)]
std_data <- apply(clean_data,2,scale)
mortal_var <- std_data[,c("Lifeexpectancy","AdultMortality","infantdeaths","under-
fivedeaths","HIV/AIDS")]
health_var <-
std_data[,c("Alcohol","percentageexpenditure","HepatitisB","Measles","BMI","Polio",
            "Totalexpenditure","Diphtheria","GDP","Population","thinness1-19years",
            "thinness5-9years","Incomecompositionofresources","Schooling")]
ccv <- cc(mortal_var,health_var)
cor <- ccv$cor
cor
model <- cancor(mortal_var,health_var)
model
Wilks(model)
sig_cor <- cor[-5]
sig_cor
squared <- sig_cor^2
data.frame(canonical_correlation = sig_cor,
           Squared_correlation = squared)
xnames <- c("Mortality_1","Mortality_2","Mortality_3","Mortality_4")
ynames <- c("H&S_1","H&S_2","H&S_3","H&S_4")
ccv$xcoef[,-5] %>% as.data.frame() %>%
  setNames(xnames)
ccv$ycoef[,-5] %>% as.data.frame() %>%
  setNames(ynames)
ccv$scores$corr.X.xscores[,-5]
ccv$scores$corr.Y.xscores[,-5]
ccv$scores$corr.X.yscores[,-5]
ccv$scores$corr.Y.yscores[,-5]

```