Canonical Correlation Analysis - Chemical Reaction Example

Stat 577 Fall 2023

11/15/2023

## Example : Chemical Reaction

The results of a planned experiment involving a chemical reaction are given. The input variables are = temperature, = concentration, = time. The yield variables are = percentage of unchanged starting material, = percentage converted to the desired product, = percentage of unwanted by-product. This data is saved in the Canvas site as ‘ChemicalReaction’.

In this problem, we wish to perform the followings:

1. find the canonical correlation values;
2. write down the canonical variate pairs, and interpret one or two canonical coefficients;
3. find the squared canonical correlation values, and interpret them in context;
4. find number of canonical variate pairs we would consider in this study based on the squared canonical correlation values;
5. report each F test statistic value based on Wilk’s lambda, its corresponding p value, and conclusion for testing hypothesis regarding canonical correlations.

## Solution

## Importing and Reading the data

setwd("D:/LianZuo/Applied Statistics Course Materials/STAT 577 - Applied Multivariate Statistics/data-577")  
CR=read.table("ChemicalReaction.dat", header=F)  
colnames(CR)=c("Obs", "start.mat", "conver.mat", "byproduct", "temp", "concentr", "time")  
print(head(CR,3), row.names = F)#row.names=F suppresses printing of labels or identifies for each row

Obs start.mat conver.mat byproduct temp concentr time  
 1 41.5 45.9 11.2 162 23 3  
 2 33.8 53.3 11.2 162 23 8  
 3 27.7 57.5 12.7 162 30 5

dim(CR)

[1] 19 7

Below we create the X-set and Y-set of variables. X set will have the input variables (“temp”, “concentr”, “time”), and Y set will have output variables (“start.mat”, “conver.mat”, “byproduct”):

ydata=CR[,2:4]  
xdata=CR[,5:7]

In order to find the canonical correlations, squared canonical correlations, canonical coefficients for canonical variates, we use cc() function from CCA package.

#install.packages("CCA")  
library(CCA)  
cca.cr=cc(xdata, ydata)  
cca.cr

$cor  
[1] 0.98152665 0.30199414 0.05733448  
  
$names  
$names$Xnames  
[1] "temp" "concentr" "time"   
  
$names$Ynames  
[1] "start.mat" "conver.mat" "byproduct"   
  
$names$ind.names  
 [1] "1" "2" "3" "4" "5" "6" "7" "8" "9" "10" "11" "12" "13" "14" "15"  
[16] "16" "17" "18" "19"  
  
  
$xcoef  
 [,1] [,2] [,3]  
temp -0.1606168 -0.0694400 -0.04909375  
concentr -0.1486092 -0.1215865 0.19117914  
time -0.2156820 0.5839192 0.03768838  
  
$ycoef  
 [,1] [,2] [,3]  
start.mat 0.17079417 0.6259417 0.3773218  
conver.mat 0.06909743 0.7299873 0.2172993  
byproduct 0.08582542 0.7127437 0.5389956  
  
$scores  
$scores$xscores  
 [,1] [,2] [,3]  
 [1,] 2.230504539 -1.16584543 -0.67087105  
 [2,] 1.152094308 1.75375064 -0.48242915  
 [3,] 0.758876244 -0.84911235 0.74275969  
 [4,] 0.111830106 0.90264530 0.85582484  
 [5,] -0.104246314 -0.93557992 -0.70407353  
 [6,] -0.751292453 0.81617773 -0.59100839  
 [7,] -0.847292173 -1.54351231 0.25182217  
 [8,] -1.494338312 0.20824534 0.36488731  
 [9,] 0.003791896 -0.01646731 0.07587565  
[10,] -1.602376522 -0.71086727 -0.41506187  
[11,] 1.609960313 0.67793265 0.56681317  
[12,] -0.739253963 -0.62439970 1.03177136  
[13,] 0.746837755 0.59146508 -0.88002005  
[14,] -0.643254243 1.73529034 0.18894079  
[15,] 0.650838034 -1.76822496 -0.03718949  
[16,] -0.487807733 0.20103131 -1.84890543  
[17,] -0.487807733 0.20103131 -1.84890543  
[18,] -0.053531875 0.26321977 1.69988471  
[19,] -0.053531875 0.26321977 1.69988471  
  
$scores$yscores  
 [,1] [,2] [,3]  
 [1,] 2.09778362 -1.1041111 0.611132567  
 [2,] 1.29398950 -0.5219566 -0.686230525  
 [3,] 0.67109241 -0.2051390 -1.266743130  
 [4,] 0.01937793 -0.6597517 -1.469499523  
 [5,] -0.14651111 -0.3299210 -1.649740910  
 [6,] -0.61377318 -0.7334426 -0.614024293  
 [7,] -0.88747010 -0.6938741 -0.516054185  
 [8,] -1.50650496 -0.5426948 2.432288140  
 [9,] -0.06693847 0.2285540 0.068736060  
[10,] -1.55848793 -2.1700100 -0.004206386  
[11,] 1.80117275 -0.3206936 1.243361361  
[12,] -0.18408905 0.3482939 0.150233450  
[13,] 0.73985934 1.0849391 0.680133375  
[14,] -0.95917850 0.5800591 0.567780019  
[15,] 0.51920558 -0.4512992 1.216491428  
[16,] -0.51895573 1.2512678 -0.064387766  
[17,] -0.40410523 1.6144702 -0.497884842  
[18,] -0.36477628 0.9335076 -0.109333771  
[19,] 0.06830943 1.6918019 -0.092051069  
  
$scores$corr.X.xscores  
 [,1] [,2] [,3]  
temp -0.7002371 -0.2147010 -0.6808608  
concentr -0.2303823 -0.1471813 0.9619052  
time -0.4415774 0.8719836 0.2113147  
  
$scores$corr.Y.xscores  
 [,1] [,2] [,3]  
start.mat 0.9770283 -0.02342752 -0.00320610  
conver.mat -0.5921615 0.15884355 -0.03437028  
byproduct -0.8455922 -0.02060390 0.02884731  
  
$scores$corr.X.yscores  
 [,1] [,2] [,3]  
temp -0.6873014 -0.06483844 -0.03903680  
concentr -0.2261263 -0.04444788 0.05515033  
time -0.4334200 0.26333395 0.01211562  
  
$scores$corr.Y.yscores  
 [,1] [,2] [,3]  
start.mat 0.9954170 -0.07757607 -0.05591924  
conver.mat -0.6033066 0.52598224 -0.59946959  
byproduct -0.8615072 -0.06822615 0.50314074

(rho.cr=cca.cr$cor)#to get canonical correlation values

[1] 0.98152665 0.30199414 0.05733448

(squared.rho.cr=cca.cr$cor^2)#to get squared canonical correlation values

[1] 0.963394567 0.091200458 0.003287242

(x.coef=cca.cr$xcoef)#to get coefficients for canonical variate for x vars

[,1] [,2] [,3]  
temp -0.1606168 -0.0694400 -0.04909375  
concentr -0.1486092 -0.1215865 0.19117914  
time -0.2156820 0.5839192 0.03768838

(y.coef=cca.cr$ycoef)#to get coefficients for canonical variate for y vars

[,1] [,2] [,3]  
start.mat 0.17079417 0.6259417 0.3773218  
conver.mat 0.06909743 0.7299873 0.2172993  
byproduct 0.08582542 0.7127437 0.5389956

Alternately, you can find the canonical correlation between the first canonical variate pair by calculating the correlation between x scores and y scores of first canonical variate pair:

cor(cca.cr$scores$xscores[,1], cca.cr$scores$yscores[,1])

[1] 0.9815267

To check if the correlation between , , etc. are zero or not, we write

cor(cca.cr$scores$xscores[,1],cca.cr$scores$yscores[,2])

[1] -1.366156e-15

cor(cca.cr$scores$xscores[,2],cca.cr$scores$yscores[,1])

[1] 1.599556e-17

To check if , we write

var(cca.cr$scores$xscores[,1])

[1] 1

var(cca.cr$scores$yscores[,1])

[1] 1

## Answer to question 1

The canonical correlation values are found to be

, , and .

## Answer to question 2

The canonical covariate pairs are , and , where

The interpretation of the canonical coefficients follows in a way that is similar to the interpretation of the coefficients of the linear regression models. For instance, consider the input variables. Suppose that we wanted an interpretation of the influence of temperature on the first canonical variate, the interpretation would be as follows:

A one unit increase in temperature would result in a decrease of units in the value of the first canonical variate for the input variables, when the other variables are held constant.

Similarly for output variables, a one unit increase in byproduct would result in an increase of units in the value of the third canonical variate for the output variables, when the other variables are held constant.

## Answer to question #3

The squared canonical correlation values are found to be

,

, and

.

We see that 96.3% of the variation in is explained by and vice versa. Only 9.12% of the variation in is explained by and vice versa, and only 0.328% of the variation in is explained by and vice versa.

## Answer to question 4

Since the first squared canonical correlation value is very high (which is 0.963394567), this implies that only the first pair of canonical variate is sufficient to explain the variation in the X and Y sets and important in the study.

## Answer to question 5

To answer Question #5, we test hypotheses regarding significance of each canonical correlation using p.asym() function from CCP package.

This function runs asymptotic tests to assign the statistical significance of canonical correlation coefficients. F-approximations of Wilks’ Lambda, the Hotelling-Lawley Trace, the Pillai-Bartlett Trace, or of Roy’s Largest Root can be used as a test statistic.

The p.asym(() has the following arguments:

rho, N, p, q, tstat = “”

rho: vector containing the canonical correlation coefficients,

N: number of observations for each variable,

p: number of independent variable,

q: number of dependent variable,

tstat: test statistic to be used. One of “Wilks” (default), “Hotelling”, “Pillai”, or “Roy”.

#install.packages("CCP")  
library(CCP)  
  
n1=nrow(CR)  
p1=ncol(xdata)  
q1=ncol(ydata)  
p.asym(rho.cr,n1,p1,q1,tstat="Wilks")

Wilks' Lambda, using F-approximation (Rao's F):  
 stat approx df1 df2 p.value  
1 to 3: 0.03315764 10.78695611 9 31.78919 1.883914e-07  
2 to 3: 0.90581210 0.35493742 4 28.00000 8.383815e-01  
3 to 3: 0.99671276 0.04947126 1 15.00000 8.269860e-01

p.asym(rho.cr,n1,p1,q1,tstat="Hotelling")

Hotelling-Lawley Trace, using F-approximation:  
 stat approx df1 df2 p.value  
1 to 3: 26.421999975 34.25074071 9 35 9.103829e-15  
2 to 3: 0.103650752 0.35414007 4 41 8.396454e-01  
3 to 3: 0.003298084 0.05166998 1 47 8.211683e-01

p.asym(rho.cr,n1,p1,q1,tstat="Pillai")

Pillai-Bartlett Trace, using F-approximation:  
 stat approx df1 df2 p.value  
1 to 3: 1.057882268 2.72352765 9 45 0.01262562  
2 to 3: 0.094487701 0.41463193 4 51 0.79729324  
3 to 3: 0.003287242 0.06252612 1 57 0.80344600

p.asym(rho.cr,n1,p1,q1,tstat="Roy")

Roy's Largest Root, using F-approximation:  
 stat approx df1 df2 p.value  
1 to 1: 0.9633946 131.5917 3 15 5.373324e-11  
  
 F statistic for Roy's Greatest Root is an upper bound.

Here we are testing three hypotheses, namely

We have summarized the hypotheses results based on Wilk’s lambda statistic in the table below:

Based on the p-value associated with the test , we reject the null hypothesis at and since the canonical correlations are ordered from largest to smallest, we concluded that at least . Therefore, first canonical variate pair is correlated.

To test whether the second or third canonical variate pairs are correlated, we perform . We found a p-value of 0.8384. From this test we can conclude that the second and third canonical variate pairs are uncorrelated, , we would stop here.

## Notes:

1. The number of canonical correlation and canonical variate pair is the same as the count of variables in the smaller set.
2. Usage of asymptotic approximations regarding the tests for canonical correlation values requires multivariate normality of the variables, or a large number of observations.
3. The hypothesis testing results regarding canonical correlation values may vary slightly depending on the test statistic implemented.