lab4

Maquieira Mariani

12/12/2019

R Markdown

```
data=read.table("P10-16.DAT")
data
```

```
##         V1       V2       V3     V4      V5
## 1 1106.000  396.700  108.400  0.787  26.230
## 2  396.700 2382.000 1143.000 -0.214 -23.960
## 3  108.400 1143.000 2136.000  2.189 -20.840
## 4    0.787   -0.214    2.189  0.016   0.216
## 5   26.230  -23.960  -20.840  0.216  70.560
```

First group is "Glucose intolerance", "Insulin response to oral glucose" and "Insulin resistance". Second group is "Relative weight" and "Fasting plasma glucose".

Variance of "Glucose intolerance", "Insulin response to oral glucose" and "Insulin resistance". (Upper left matrix)

```
v1= as.matrix(data[1:3,1:3])
v1
```

```
##       V1     V2     V3
## 1 1106.0  396.7  108.4
## 2  396.7 2382.0 1143.0
## 3  108.4 1143.0 2136.0
```

Variance of "Relative weight" and "Fasting plasma glucose".

```
v2=as.matrix(data[4:5,4:5])
v2
```

```
##      V4     V5
## 4 0.016  0.216
## 5 0.216 70.560
```

Covariance 1st and 2nd group

```
v12=as.matrix(data[1:3,4:5])
v21=as.matrix(data[4:5,1:3])
```

a) Test at the 5% level if there is any association between the groups of variables.

It is often linear combinations of variables that are interesting and useful for predictive or comparative purposes. The main task of canonical correlation analysis is to summarize the associations between the groups.

When $\sum_{12} = 0$ (v12=0) all the canonical correlations must be zero, and there is no point in pursuing a canonical correlation analysis. Therefore we need to test $\sum_{12} = 0$.

As stated in 10-39 formula in the book:

Reject H_0: $\mathbf{\Sigma}_{12} = \mathbf{0}$ ($\rho_1^* = \rho_2^* = \cdots = \rho_p^* = 0$) at significance level α if

$$-\left(n - 1 - \frac{1}{2}(p + q + 1)\right) \ln \prod_{i=1}^{p} (1 - \widehat{\rho}_i^{*2}) > \chi_{pq}^2(\alpha) \qquad (10\text{-}39)$$

where $\chi_{pq}^2(\alpha)$ is the upper (100α)th percentile of a chi-square distribution with pq d.f.

```
cm=as.matrix(data)
n=46
p=3
q=2
qchisq(1-0.05,df=p*q)
```

```
## [1] 12.59159
```

If our result is greater than 12.59, we shall accept the null hypothesis.

```
p1=-(n-1-1/2*(p+q+1))
p2=log(det(v1)*det(v2)/det(cm))

p1*p2
```

```
## [1] -13.74948
```

We reject the null hypotehsis. Since the null hypothesis is rejected, it is natural to examine the "significance" of the individual canonical correlations.

b) How many pairs of canonical variates are significant?

Since the canonical correlations are ordered from the largest to the smallest, we can begin by assuming that the first canonical correlation is nonzero and the remaining canonical correlations are zero. If this hypothesis is rejected, we assume that the first two canonical correlations are nonzero, but the remaining canonical correlations are zero, and so forth.

We assume that the first canonical correlation is non-zero and the remaining are zero!

From example 10.1 from book:

From Result 10.1, $\mathbf{f}_1 \propto \boldsymbol{\rho}_{22}^{-1/2}\boldsymbol{\rho}_{21}\boldsymbol{\rho}_{11}^{-1/2}\mathbf{e}_1$ and $\mathbf{b}_1 = \boldsymbol{\rho}_{22}^{-1/2}\mathbf{f}_1$. Consequently,

$$\mathbf{b}_1 \propto \boldsymbol{\rho}_{22}^{-1}\boldsymbol{\rho}_{21}\mathbf{a}_1 = \begin{bmatrix} .3959 & .2292 \\ .5209 & .3542 \end{bmatrix}\begin{bmatrix} .8561 \\ .2776 \end{bmatrix} = \begin{bmatrix} .4026 \\ .5443 \end{bmatrix}$$

We must scale $\mathbf{b}_1$ so that

$$\text{Var}(V_1) = \text{Var}(\mathbf{b}_1'\mathbf{Z}^{(2)}) = \mathbf{b}_1'\boldsymbol{\rho}_{22}\mathbf{b}_1 = 1$$

The vector $[.4026, .5443]'$ gives

$$[.4026, .5443]\begin{bmatrix} 1.0 & .2 \\ .2 & 1.0 \end{bmatrix}\begin{bmatrix} .4026 \\ .5443 \end{bmatrix} = .5460$$

Using $\sqrt{.5460} = .7389$, we take

$$\mathbf{b}_1 = \frac{1}{.7389}\begin{bmatrix} .4026 \\ .5443 \end{bmatrix} = \begin{bmatrix} .5448 \\ .7366 \end{bmatrix}$$

The first pair of canonical variates is

$$U_1 = \mathbf{a}_1'\mathbf{Z}^{(1)} = .86Z_1^{(1)} + .28Z_2^{(1)}$$
$$V_1 = \mathbf{b}_1'\mathbf{Z}^{(2)} = .54Z_1^{(2)} + .74Z_2^{(2)}$$

and their canonical correlation is

$$\rho_1^* = \sqrt{\rho_1^{*2}} = \sqrt{.5458} = .74$$

This is the largest correlation possible between linear combinations of variables from the $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$ sets.

The second canonical correlation, $\rho_2^* = \sqrt{.0009} = .03$, is very small, and consequently, the second pair of canonical variates, although uncorrelated with members of the first pair, conveys very little information about the association between sets. (The calculation of the second pair of canonical variates is considered in Exercise 10.5.)

We note that U_1 and V_1, apart from a scale change, are not much different from the pair

$$\widetilde{U}_1 = \mathbf{a}'\mathbf{Z}^{(1)} = [3, 1]\begin{bmatrix} Z_1^{(1)} \\ Z_2^{(1)} \end{bmatrix} = 3Z_1^{(1)} + Z_2^{(1)}$$
$$\widetilde{V}_1 = \mathbf{b}'\mathbf{Z}^{(2)} = [1, 1]\begin{bmatrix} Z_1^{(2)} \\ Z_2^{(2)} \end{bmatrix} = Z_1^{(2)} + Z_2^{(2)}$$

```
library(expm) #sqrtm does matrix square root.
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'expm'
```

```
## The following object is masked from 'package:Matrix':
##
##     expm
```

```r
form=solve(sqrtm(v1)) %*% v12 %*% solve(v2) %*% v21 %*% solve(sqrtm(v1))
eig=eigen(form)
sqrt(eig$values)
```

```
## Warning in sqrt(eig$values): NaNs produced
```

```
## [1] 0.5173449 0.1255082       NaN
```

This is the largest correlation possible between linear combinations of variables from the 2 groups.

The third canonical correlation is very small to calculate, and consequently, the third pair of canonical variates, although uncorrelated with members of the first pair, conveys very little information about the association between sets.

c) Interpret the significant squared canonical correlations.

From book: because of its multiple correlation coefficient interpretation, the squared canonical correlation is the proportion of the variance of canonical variate "explained" by the set. It is also the proportion of the variance of canonical variate "explained" by the set Therefore, it is often called the shared variance between the two sets. The largest value, is sometimes regarded as a measure of set "overlap."

d) Interpret the canonical variates by using the coefficients and suitable correlations.

from example 10.1

```r
solve(sqrtm(v1))%*%eig$vectors[,1]
```

```
##             [,1]
## [1,]  0.01310065
## [2,] -0.01443825
## [3,]  0.02339972
```

$$\hat{U_1} = 0.013z_1^{(1)} - 0.014z_2^{(1)} + 0.023z_3^{(1)}$$

```r
form2=solve(sqrtm(v2)) %*% v21 %*% solve(v1) %*% v12 %*% solve(sqrtm(v2))
eig2=eigen(form2)
solve(sqrtm(v2))%*%eig2$vectors[,1]
```

```
##             [,1]
## [1,] -8.06557508
## [2,]  0.01915905
```

$$\hat{V_1} = -8.066z_1^{(2)} + 0.0191z_2^{(2)}$$

Since each coefficient describes weights. We can say the group U, is mostly represented by 'insulin resistance'(weight 0.023)

Group V is pretty much represented by 'relative weight'

e) Are the "significant" canonical variates good summary measures of the respective data sets?

Following example 10.7

```
a=solve(sqrtm(v1))%*%eig$vectors[,1]
b=solve(sqrtm(v2))%*%eig2$vectors[,1]

1/3*(a[1]^2+a[2]^2+a[3]^2)
```

```
## [1] 0.0003092125
```

```
1/2*(b[1]^2+b[2]^2)
```

```
## [1] 32.52693
```

The first sample canonical variate of the first group accounts .0003% of the set's total sample variance.

The first sample canonical variate of the second group explains 32.52% of the set's total sample variance.

We might thus infer that the first sample linear combination of the second group is a "better" representative of its set than the linear combination of the first group is is of its set.

f) Give your opinion on the success of this canonical correlation analysis.

There is no definition for 'success'. But we can say that since the canonical correlation analysis seeks to identify and quantify the associations between two sets of variables, and we have identified two of such, with correlation 0.5173449 and 0.1255082 respectively, we consider this a success.