# Applied Multivariate Analysis (567): Report 1

By Yujing Ma

## Dataset

The data used for this report comes from the sixth wave of World Values Survey of Philippines, between 2010 and 2014 (1200 observations and 257 variables)

## Problems

**Q1** - Using variables V4,V5,V6,V7,V8,V9, from your assigned country, do a Hoteling's $T2$ test to test the null hypothesis $H0$:$\mu 0$ = [1.1, 1.5, 1.69, 2.43, 1.95, 2.02] with an $\alpha$=0.05. The order in $\mu 0$ corresponds to the variables V4-V9 for the United States.

1. Estimate and show the sample means.

```
> samplemeans = colMeans(q1data[5:10],na.rm=TRUE)
> round(samplemeans,2)
 V4   V5   V6   V7   V8   V9
1.01 1.84 2.47 2.29 1.11 1.16
```

2. Estimate and show the variance-covariance matrix

```
> S =cov(q1data[5:10],use = "complete.obs")#missing value
> round(S,2)
      V4   V5    V6   V7   V8   V9
V4  0.02 0.01 -0.01 0.00 0.00 0.01
V5  0.01 0.44  0.14 0.08 0.00 0.02
V6 -0.01 0.14  0.74 0.18 0.01 0.03
V7  0.00 0.08  0.18 1.00 0.05 0.04
V8  0.00 0.00  0.01 0.05 0.15 0.03
V9  0.01 0.02  0.03 0.04 0.03 0.17
```

3. Estimate and show the inverse of the variance-covariance matrix.

```
> S.inv = solve(S)
> round(S.inv,2)
      V4    V5    V6    V7    V8    V9
V4 58.28 -1.32  1.25 -0.04  0.01 -3.77
V5 -1.32  2.49 -0.48 -0.11  0.12 -0.14
V6  1.25 -0.48  1.53 -0.23 -0.01 -0.22
V7 -0.04 -0.11 -0.23  1.07 -0.34 -0.16
V8  0.01  0.12 -0.01 -0.34  7.12 -1.16
V9 -3.77 -0.14 -0.22 -0.16 -1.16  6.30
```

4. Estimate and show the critical value.

Critical value is following the F-distribution with parameters p and n-p, and n is the sample size and p the number of variables. Since $\alpha$=0.05, and n=1200, p=6, the critical values is 12.69

```
> critical = (((n - 1)*p)/(n - p))*qf(1-alpha,p,n-p)
[1] 12.69
```

5. Estimate and show the $T2$.

$T^2$ =10877.26

```
> #Hotelling's T2
> T2 = n*t(samplemeans - mu0)%*%S.inv%*%(samplemeans - mu0)
          [,1]
[1,] 10877.26
```

6. Do you reject or fail to reject the null?

Reject the null hypothesis.
In this test, $H_0$:$\mu_0$ = [1.1, 1.5, 1.69, 2.43, 1.95, 2.02] with $\alpha$=0.05
$T^2$ is 10877.26, the critical values is 12.69. We reject the null hypothesis if $T^2$>critical.
Therefore, we reject the null at the 5% level of significance.

7. Are the values $\mu$0 plausible mean values of the data?

The values $\mu$0 are not plausible mean values of the data.
Based on the result of the Hoteling's $T^2$ test, this data does not support one or more of the hypothesized values. Variables V4-V9 for Philippines are totally different from the United States.

8. Estimate and show the simultaneous $T^2$ confidence intervals per variable.

```
> t2CI = data.frame(var =  c("Family","Friend","Leisure time","Politics","Work","Religion"),
 mu0 = c(1.1,1.5,1.69,2.43,1.95,2.02), lower=c(v4low,v5low,v6low,v7low,v8low,v9low), upper=c
(v4up,v5up,v6up,v7up,v8up,v9up))
> t2CI$plausible = ifelse(t2CI$mu0>t2CI$lower & t2CI$mu0<t2CI$upper, "Yes","No")
> t2CI
             var  mu0 lower upper plausible
V4        Family 1.10  1.00  1.03        No
V5        Friend 1.50  1.78  1.91        No
V6 Leisure time 1.69  2.38  2.56        No
V7      Politics 2.43  2.19  2.40        No
V8          Work 1.95  1.07  1.15        No
V9      Religion 2.02  1.12  1.20        No
```

9. Estimate and show the Bonferoni confidence intervals per variable.

```
> BonCI = data.frame(var =  c("Family","Friend","Leisure time","Politics","Work","Religion
"),mu0 = c(1.1,1.5,1.69,2.43,1.95,2.02),lower=c(v4low1,v5low1,v6low1,v7low1,v8low1,v9low1),u
pper=c(v4up1,v5up1,v6up1,v7up1,v8up1,v9up1))
> BonCI$plausible = ifelse(BonCI$mu0>BonCI$lower & BonCI$mu0<BonCI$upper, "Yes","No")
> BonCI
             var  mu0 lower upper plausible
V4        Family 1.10  1.01  1.02        No
V5        Friend 1.50  1.81  1.88        No
V6 Leisure time 1.69  2.43  2.51        No
V7      Politics 2.43  2.25  2.34        No
V8          Work 1.95  1.09  1.13        No
V9      Religion 2.02  1.14  1.18        No
```

10. Individually, which variables are plausible, which ones are not?

$H_0$:$\mu_0$ = [1.1, 1.5, 1.69, 2.43, 1.95, 2.02]. According to the simultaneous $T$2 confidence intervals and the Bonferoni confidence intervals, all the variables are not plausible.

| Variable | $T$2 | | Bonferoni | |
|---|---|---|---|---|
| | **Confident interval** | **Plausible** | **Confident interval** | **Plausible** |
| V4 | (1.00, 1.03) | No | (1.01, 1,02) | No |
| V5 | (1.78, 1.91) | No | (1.81, 1.88) | No |
| V6 | (2.38, 2.56) | No | (2.43, 2.51) | No |
| V7 | (2.19, 2.40) | No | (2.25, 2.34) | No |

| | | | | |
|---|---|---|---|---|
| V8 | (1.07, 1,15) | No | (1.09, 1.13) | No |
| V9 | (1.12, 1.20) | No | (1.14, 1.18) | No |

11. What do you conclude?

Based on the results, the people of Philippines and the United States have different values.

Philippines believes family is the most important, work and religion are very important, friends and policies are rather important, and leisure time are not very important. While Americans regard the family and friend are very important, leisure time, work and religion are rather important, and policies are not very important.

This may causes by the development and culture of these countries, since Philippines is a developing country and it has been heavily influenced by both Asian and Western cultures. This is total different from those of the United States.

**Q2** - Using the variables from the previous question:

1. Estimate and show the simultaneous confidence intervals using the process for large samples.

```
> q2 = data.frame(var = c("Family","Friend","Leisure time","Politics","Work","Religion"),mu0
 = c(1.1,1.5,1.69,2.43,1.95,2.02),lower=c(v4low2,v5low2,v6low2,v7low2,v8low2,v9low2),upper=c
(v4up2,v5up2,v6up2,v7up2,v8up2,v9up2))
> q2$plausible = ifelse(q2$mu0>q2$lower & q2$mu0<q2$upper, "Yes","No")
> q2
             var  mu0    lower    upper plausible
v4        Family 1.10 1.001147 1.028853        No
v5        Friend 1.50 1.776248 1.911825        No
v6 Leisure time 1.69 2.380704 2.556640        No
v7      Politics 2.43 2.189876 2.395204        No
v8          Work 1.95 1.069854 1.148661        No
v9      Religion 2.02 1.116489 1.201844        No
```

2. Compare to the $T2$ and Bonferroni CIs estimated in the previous question. Do your conclusions still hold? Discuss.

The conclusions of question 1 still hold. The reasons are as following:

The critical values of the three methods are similar. In addition, for this large sample size n=1200, the confident intervals differences are typically different in the thousandths, or digits after that. Therefore, the confidence intervals from the process for large samples still holds the conclusion.

For $H_0$:$\mu_0$ = [1.1, 1.5, 1.69, 2.43, 1.95, 2.02] with $\alpha$=0.05

| **Methods** | **Large sample** | ***T2* CI** | **Bonferroni CI** |
|---|---|---|---|
| **Critical value** | 12.59 | 12.70 | 2.64 |

**Q3** Variable Y002 in the WVS dataset is a post-materialist index with three categories: materialist (1), mixed (2) and post-materialist (3). Using data from your randomly assigned country perform a MANOVA where Y002 defines the population and the Xs are: V96 and V97

1. Estimate and show the treatment (between) sum of squares and cross-products and its degree of freedom.

```
> B = as.matrix(q3[c("te1","te2")])
> B = t(B)%*%B
```

```
      te1   te2
te1 19.24 12.41
te2 12.41 15.42
> Bdf = g-1
 [1] 2
```

| source | Matrix of sum of squares and cross products | Degree of freedom |
|---|---|---|
| **treatment** | $\begin{pmatrix} 19.24 & 12.41 \\ 12.41 & 15.42 \end{pmatrix}$ | 2 |

2. Estimate and show the residual (within) sum of squares and cross-products and its degree of freedom.

```
> W = as.matrix(q3[c("residuals1","residuals2")])
> W = t(W)%*%W
           residuals1 residuals2
residuals1   11428.11    2539.10
residuals2    2539.10   11160.12
Its degrees of freedom is 1178.
> Wdf = n-g
 [1] 1178
```

| source | Matrix of sum of squares and cross products | Degree of freedom |
|---|---|---|
| **residual** | $\begin{pmatrix} 11428.11 & 2539.10 \\ 2539.10 & 11160.12 \end{pmatrix}$ | 1178 |

3. Estimate and show the total (between/within) sum of squares and cross-products and its degrees of freedom.

```
> BW = as.matrix(q3[c("dev1","dev2")])
> BW = t(BW)%*%BW
         dev1     dev2
dev1 11447.36  2551.51
dev2  2551.51 11175.55
> BWdf = n-1
 [1] 1180
```

| source | Matrix of sum of squares and cross products | Degree of freedom |
|---|---|---|
| **residual** | $\begin{pmatrix} 11447.36 & 2551.51 \\ 2551.51 & 11175.55 \end{pmatrix}$ | 1180 |

4. Estimate and show the Wilks' lambda.

The Wilks' lambda $\Lambda^*=1.00$

```
> lambda = det(W)/det(BW)
 [1] 0.9972976
> round(lambda,2)
[1] 1
```

5. Estimate and show the F-statistic (use the formula shown in the slides).

```
> F = ((1-sqrt(lambda))/sqrt(lambda))*((n-g-1)/(g-1))
 [1] 0.80
```

6. Estimate and show the critical value.

```
> critical = qf(1-alpha,2*(g-1),2*(n-g-1))
 [1] 2.38
```

7. What do you conclude?

We can't eject the null hypothesis, and the variance-covariance matrices across groups are equal. In this test, with $\alpha$=0.05, T=0.80<F2, 1178(0.05) =2.38, the p-values is 0.53. We cannot reject the null hypothesis.

```
> pval = 1 - pf(F,2*(g-1),2*(n-g-1))
[1] 0.53
```

In Philippines, the review on income equality and private vs state ownership of business would affect i ts Post-materialist index equally.

**Q4** It is believed that variables V131, V132, V133, V134, V135, V136, V137, V138, and V139 can be grouped into one component-variable that captures, at least, two thirds of the common variance. Is this possible using the data from your assigned country? If not, how many component-variables do we need to achieve that goal?

1. Estimate and show the eigenvalues

```
> eigenvalues = eigen(rho)$values
> round(eigenvalues,2)
[1] 3.14 1.12 0.87 0.81 0.76 0.63 0.58 0.55 0.53
```

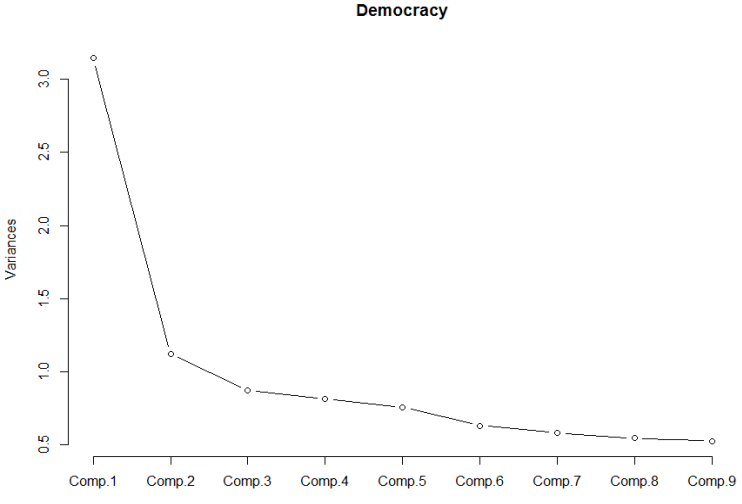
2. Estimate and show the percent of variance each eighenvalue explains.

```
> round( eigenvalues/sum(eigenvalues) * 100, 2) # % Variance
[1] 34.92 12.46  9.69  9.04  8.44  7.03  6.48  6.07  5.87
```

3. Estimate and show the cumulative variance.

```
> round( cumsum(eigenvalues)/sum(eigenvalues) * 100, 2)
[1] 34.92  47.38  57.07  66.12  74.55  81.58  88.06  94.13 100.00
```

4. Plot and show a scree plot.




An elbow occurs in the plot at about i=2, and all the eigenvalues after that $\hat{\lambda}2$ are all relatively small and about the same size. So, the number of principal components is 2.

5. What do you conclude?

The first principal component explains 34.92% of the total variance. The first two principal component explains 47.38% of the total variance. Sample variation is not summarized well by only these two components. In addition, they're related to all the variables, so no variable reduction. So the results of those 9 survey questions could not describe the review of democracy in Philippines.

**Q5** Researches believed that two uncorrelated latent variables could be created with V96, V97, V98, V99, V100, and V101. Is this possible? Run a factor analysis using a principal-component/principal factor method to answer the question. Fill in the following table, show your work in estimating, step-by-step, the values. What do you conclude?

1. Eigenvalues and estimate factor loadings

```
> eigenvalues = eigen(R)$values
> round(eigenvalues,2)
```

```
[1] 1.60 1.23 0.97 0.86 0.75 0.59
> f1 = round ( sqrt(eigenvalues[1])*eigenvectors[,1], 2)
 [1] -0.46 -0.48 -0.47 -0.64 -0.63 -0.35
> f2 = round ( sqrt(eigenvalues[2])*eigenvectors[,2], 2)
 [1]  0.61  0.45  0.34 -0.53 -0.51  0.01
```

2. Communalities

```
> h2 = round ( f1^2 + f2^2, 2)
 [1] 0.58 0.43 0.34 0.69 0.66 0.12
```

3. Specific variances (uniqueness) and cumulative proportion of total sample variance explained

```
> psi = 1 -h2
 [1] 0.42 0.57 0.66 0.31 0.34 0.88
> CPTSSC=round( cumsum(eigenvalues)/sum(eigenvalues) * 100, 2) # Cum. variance
 [1]  26.67  47.10  63.26  77.66  90.17 100.00
```

4. Table

| Variable | f1 | f2 | h2 | psi |
|---|---|---|---|---|
| V96 | -0.46 | 0.61 | 0.58 | 0.42 |
| V97 | -0.48 | 0.45 | 0.43 | 0.57 |
| V98 | -0.47 | 0.34 | 0.34 | 0.66 |
| V99 | -0.64 | -0.53 | 0.69 | 0.31 |
| V100 | -0.63 | -0.51 | 0.66 | 0.34 |
| V101 | -0.35 | 0.01 | 0.12 | 0.88 |
| Eigenvalues | 1.60 | 1.23 | | |
| CPTSSV | 26.67 | 47.10 | | |

5. Factor loading after rotation

```
       f1    f2
[1,]  0.06 0.76
[2,] -0.06 0.65
[3,] -0.13 0.57
[4,] -0.83 0.02
[5,] -0.81 0.03
[6,] -0.26 0.24
```

So it's impossible that two uncorrelated latent variables could be created with V96, V97, V98, V99, V100.

**Q6** Question V81 compares two statments, category 1 represent those who think that "protecting the e nvironment should be given priority, even if it causes slower economic growth and some loss of jobs", category 2 represent those who think that "economic growth and creating jobs should be the top priori ty, even if the environment suffers to some extent". Recode V81 so category 1 is 0, and category 2 is 1. Use variables V238 (treat it as discrete-continous variable), V240 (recode this variable so 1 is 'female' and 0 is 'male'), and V242 to run a discriminant analysis.

1. Assuming equal cost ratio and equal prior probability ratios, run an LDA and a QDA, show their resp ective confusion matrices and estimate the apparent error rate.

```
># 1. Linear discriminant model
> lda.pred= predict(lda1)
> lda.table= table(actual = q6data$V81,pred
icted=lda.pred$class)
> lda.table # confusion matrices
      predicted
actual   0   1
     0 390 374
     1 194 223

> #the apparent error rate
> round((374+194)/(390+374+194+223),2)
[1] 0.48
```

```
> # 2.Quadratic discriminant model
> qda.pred= predict(qda1)
> qda.table= table(actual = q6data$V81,pred
icted=qda.pred$class)
> qda.table # confusion matrices
      predicted
actual   0   1
     0 333 431
     1 147 270

> # the apparent error rate
> round((431+147)/(333+431+147+270),2)
[1] 0.49
```

2. Using the Lachenbruch's "holdout" procedure estimate the actual error

```
>#  1. Linear discriminant model-hold out
> lda2= lda(V81~ V238 + V240 + V242, data=q6
data, prior=c(0.5,0.5),CV=TRUE)
> lda2.table = table(actual = q6data$V81,pre
dicted = lda2$class)
> lda2.table  # confusion matrices
      predicted
actual   0   1
     0 378 386
     1 208 209
> #the actual error rate
> round((386+208)/(378+386+208+209),2)
[1] 0.5
```

```
>#  2.Quadratic discriminant model-hold out
> qda2= qda(V81~ V238 + V240 + V242, data=q
6data, prior=c(0.5,0.5), CV=TRUE)
> qda2.table= table(actual = q6data$V81,pre
dicted=qda2$class)
> qda2.table   # confusion matrices
      predicted
actual   0   1
     0 328 436
     1 158 259
> #the actual error rate
> round((436+158)/(328+436+158+259),2)
[1] 0.5
```

3. Compare the APER and AER estimated in 2 and 3 above. What do you conclude?

| | **LDA** | **QDA** |
|---|---|---|
| **APER (the apparent error rate)** | 0.48 | 0.49 |
| **AER (the actual error rate)** | 0.50 | 0.50 |

- The APER is underestimate the AER, since the sample size is large. Since the same data was used to construct the classification rule are also used to evaluate it.
- For large sample size, QDA tends to work well.

4. Run and show a test for equality of variance-covariances. What do you conclude?

```
> #H0: equal variances.
> boxM(cbind(q6data$V238,q6data$V240,q6data$V242),q6data$V81)
        Box's M-test for Homogeneity of Covariance Matrices
data:  cbind(q6data$V238, q6data$V240, q6data$V242)
Chi-Sq (approx.) = 8.07, df = 6, p-value = 0.23
```

For $H_0$: equal variances, the critical values is 8.07. We couldn't reject the null hypothesis since p-value=0.23.Therefore, covariance matrices are not significantly different in Philippines data.

5. What would an upper-middle class, 41 year old woman respond? How about a man?

Since the p-value of the Wald test of the logistic regression is 0.53, so we don't use the logistic regressi on to predict the response.

| | **LDA prediction** | **QDA prediction** | **Conclusion** |
|---|---|---|---|
| **Woman** | 0 1<br>1 0.52 0.48 | 0 1<br>1 0.47 0.53 | An upper-middle class, 41 year old woman more likely choose category 1 of the statements based on LDA method. But based on QDA prediction she more likely choose category 1. |
| **Man** | 0 1<br>1 0.49 0.51 | 0 1<br>1 0.48 0.52 | An upper-middle class, 41 year old man more likely choose category 2 of the statements. |

6. What would a lower-middle class, 41 year old woman respond? How about a man?

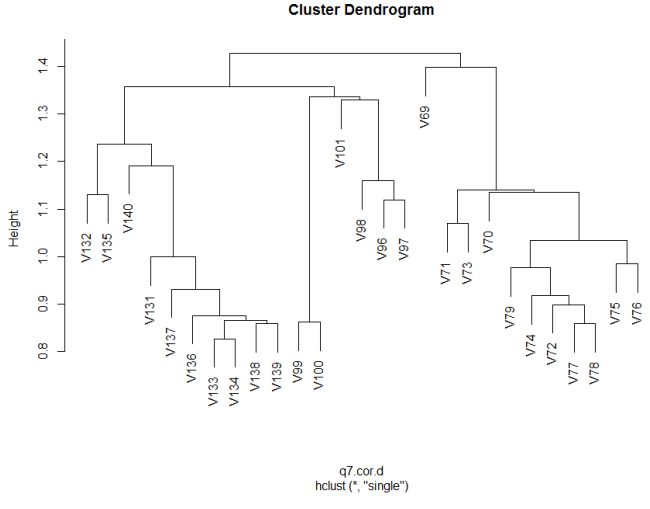
| | LDA prediction | QDA prediction | Conclusion |
|---|---|---|---|
| **Woman** | 0 1<br>1 0.51 0.49 | 0 1<br>1 0.48 0.52 | A lower-middle class, 41 year old woman more li kely choose category 1 of the statements based o n LDA method. But based on QDA prediction she more likely choose category 1. |
| **Man** | 0 1<br>1 0.49 0.51 | 0 1<br>1 0.47 0.53 | A lower-middle class, 41 year old man more likely choose category 2 of the statements. |

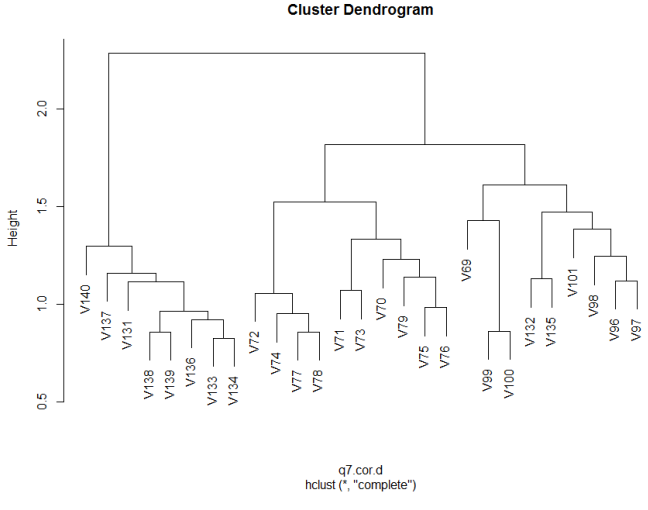
7. What do you conclude?

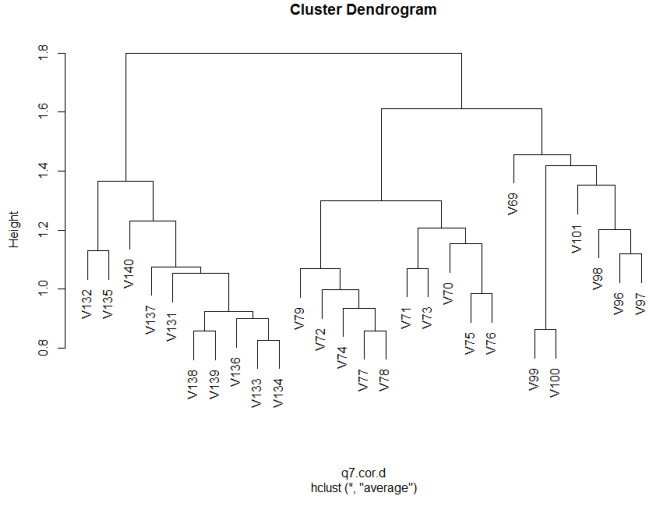
- In Philippines, no matter which the classes they are in, men more likely believe economic growth and creating jobs are more important than environment protection.
- Because Philippines is a developing country, economic growth and creating jobs are as important as protecting the environment.

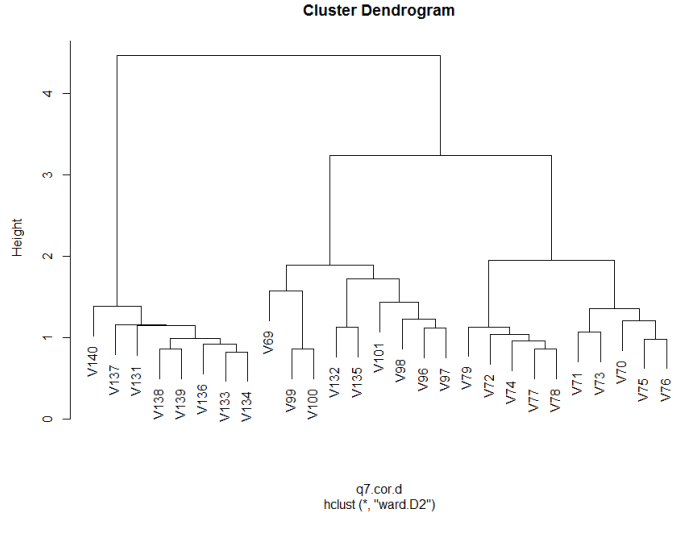
**Q7** Using variables V69-V79, V96-V101, and V131-V140. Run a cluster analysis data from your assigned country. Here the goal is to cluster variables.

1. Run single, complete, average and Ward clustering. Show the dendrograms.



2. How many clusters do you see in each method?

| | single | complete | average | Ward |
|---|---|---|---|---|
| **Number of clusters** | 4 | 5 | 4 | 7 |

3. Make a table comparing the clustering methods.

```
> q7.vars
     vars single comp ave ward
v69   v69      1    1   1    1
v70   v70      2    2   2    2
v71   v71      2    2   2    2
v72   v72      2    3   2    3
v73   v73      2    2   2    2
v74   v74      2    3   2    3
v75   v75      2    2   2    2
v76   v76      2    2   2    2
v77   v77      2    3   2    3
v78   v78      2    3   2    3
v79   v79      2    2   2    3
v96   v96      3    4   3    4
v97   v97      3    4   3    4
v98   v98      3    4   3    4
v99   v99      4    1   4    1
v100 v100      4    1   4    1
v101 v101      3    4   3    4
v131 v131      5    5   5    5
v132 v132      5    4   5    4
v133 v133      5    5   5    5
v134 v134      5    5   5    5
v135 v135      5    4   5    4
v136 v136      5    5   5    5
v137 v137      5    5   5    5
v138 v138      5    5   5    5
v139 v139      5    5   5    5
v140 v140      5    5   5    5
```

4. Any specific pattern in the grouping of variables? Discuss.

- V70, V71, V73, V75 and V76 are always in the same cluster, since people who are creative and willing to take risks want to be successful and rich as well. These variables identify the same kinds of people. In addition, V131, V133, V134 and V136-140 are always in the same cluster, because they are all related to democracy and wealth.
- Single method always tends to merge the cluster. While, complete method do not merge close groups because of outlier members that are far apart. For example, V72, V74 and V135 will be put into group 2, while complete and Ward clustering put them in anther groups.