Notes:

1. The assignment is visualized in a way to show the key points in **BOLD BLACK**, and references to more details and further works in Appendix in dark gold.
2. All the test are interpreted under 95% confidence level.

I. Could you determine, with the available information, which are the variables discriminating between being or not a born-global firm? Born Global firms are those starting their exports 5 years, as a maximum, after the date of inception (you have to create this variable considering the information in variable p09).

Is the firm born global?	Freq.	Percent	Cum.
No	27	50.94	50.94
Yes	26	49.06	100.00
Total	53	100.00	

First, we create the born global binary variable and give it labels to make it more intuitive. The frequency of born global and not born global firms is as follow: (Code in Appendix I)

Second, we run the **linear discriminant analysis with the available variables and previous factors (behaviour, performance, management_behaviour) instead p07_1:p07_13 to have a more simplified result to interpret.**

```
. xi: discrim lda behaviour performance management_behaviour p02 i.p03 p04 p05 p06 p08, group(b
> orn_global) priors(0.5, 0.5)
i.p03             _Ip03_1-3           (naturally coded; _Ip03_1 omitted)

Linear discriminant analysis
Resubstitution classification summary
```

Key
Number Percent

True born_global	Classified No	Yes	Total
No	16	2	18
	88.89	11.11	100.00
Yes	2	12	14
	14.29	85.71	100.00
Total	18	14	32
	56.25	43.75	100.00
Priors	0.5000	0.5000	

Now let's see how different the means are regarding being born global and not:

```
.
. estat grsummarize

Estimation sample discrim lda
Summarized by born_global
```

Mean	born gl~l No	Yes	Total
behaviour	-.1744844	.0336299	-.0834344
performance	.0589723	.3685208	.1943998
management~r	-.2711583	.010552	-.14791
p02	1971.444	1985.357	1977.531
_Ip03_2	.5	.5	.5
_Ip03_3	.1666667	.0714286	.125
p04	.4444444	.6428571	.53125
p05	1992.722	1987.143	1990.281
p06	17.08339	50.39714	31.65816
p08	54.61111	31.34429	44.43187
N	18	14	32

Most of the variables show different means for the groups. For example, p02 shows the year of foundation and for the born global firms the year is higher which makes sense based on trends in globalization. Moreover, the factors have interesting distinct means for the groups. Behavior mean is negative for the not-born global firms while for the born global firms its slightly positive. The same means for management behavior factor.

But is the difference in means coherent and significant? Let's carry on correlation and ANOVA.

```
. estat correlations

Pooled within-group correlation matrix
```

	behavi~r	perfor~e	manage~r	p02	_Ip03_2	_Ip03_3	p04
behaviour	1.00000						
performance	0.19316	1.00000					
management~r	0.14553	0.02617	1.00000				
p02	-0.05272	0.06090	-0.08131	1.00000			
_Ip03_2	-0.12066	-0.36942	-0.49580	0.10147	1.00000		
_Ip03_3	-0.07893	0.23619	0.21861	0.22466	-0.38188	1.00000	
p04	-0.19770	0.08713	0.04298	0.13547	-0.06388	0.00465	1.00000
p05	-0.04575	0.00497	-0.10215	0.68950	0.13507	0.16016	0.02527
p06	0.49313	0.15995	0.34548	-0.37230	-0.28843	-0.06224	0.00188
p08	-0.09098	0.10503	0.32072	-0.19781	-0.16431	0.35683	-0.40499

	p05	p06	p08
p05	1.00000		
p06	-0.42264	1.00000	
p08	-0.15924	0.13198	1.00000

High intra-group correlation shows conformity within the groups, while a lot of variables show **low intra-group correlation showing non-similarity in groups**.

Only p02 – p05 and p06 - behaviour factor show relatively higher correlations. Without high intra-correlation, the reliability of discriminated groups is doubted.

```
. estat anova

Univariate ANOVA summaries
```

Variable	Model MS	Resid MS	Total MS	R-sq	Adj. R-sq	F	Pr > F
behaviour	.34107852	26.9729	26.113809	0.0125	-0.0204	.37936	0.5426
performance	.75458464	21.941794	21.258335	0.0332	0.0010	1.0317	0.3179
managemen~r	.62496554	25.657932	24.850416	0.0238	-0.0088	.73073	0.3994
p02	1524.31	9861.6587	9592.712	0.1339	0.1050	4.6371	0.0394
p03	.28571429	13.714286	13.281106	0.0204	-0.0122	.625	0.4354
p04	.31001984	7.6587302	7.421675	0.0389	0.0069	1.2144	0.2792
p05	245.14335	3167.3254	3073.0615	0.0718	0.0409	2.3219	0.1380
p06	8739.7239	24629.556	24116.981	0.2619	0.2373	10.645	0.0028
p08	4263.0932	86781.771	84119.878	0.0468	0.0151	1.4737	0.2342

Number of obs = 32 Model df = 1 Residual df = 30

As we suspected, **only p02 and p06 have significant different means for the groups.** Now that we know only two coefficients have significant different means, can we continue with the discriminant function we derived at the beginning with all the variables? To do this, we have to check how the groups have been separated doing a canonical linear discriminant test or one-way ANOVA on discriminant scores by born global group.

```
. estat canontest

Canonical linear discriminant analysis
```

Fcn	Canon. Corr.	Eigen-value	Variance Prop.	Variance Cumul.	Like-lihood Ratio	F	df1	df2	Prob>F
1	0.7491	1.27861	1.0000	1.0000	0.4389	3.1255	9	22	0.0140 e

Ho: this and smaller canon. corr. are zero; e = exact F

estat canontest shows the canonical correlations test that relates to our LDA model. Since we have only one discriminating variable - born global - we only have one canonical correlation, eigen value, and accounted for variance. **The likelihood ratio test has an exact 0.014 p-value for the F-test (e beside the p-value shows that it's the exact value) asserting that the canonical correlation is likely not zero.**

```
Total-sample standardized canonical discriminant function coefficients
```

	function1
behaviour	.3758906
performance	-.18837
management~r	-.1480722
p02	-1.126749
_Ip03_2	-.2087542
_Ip03_3	.1287878
p04	.1526578
p05	.6946454
p06	-.9370397
p08	.2810074

The total-sample standardized canonical discriminant function coefficients are the coefficients that apply to the discriminating variables after they have been standardized by the total-sample covariance.

The standardized coefficients are better for comparison. As you can see, p05 and p02 and p06 have the highest absolute values , and behavior factor, p03 and p08 are not as important as them. Remember that p02 and p06 where the variables with significant means between the two groups.

The discriminant function coefficients are the multipliers that are used to calculate the discriminant score. The coefficients can be estimated using standardized sample or unstandardized sample as well. total-sample standardized coefficients is shown here and pooled within-group standardized, and unstandardized result are available at the appendix II.

```
.
. estat structure

Canonical structure
```

	function1
behaviour	-.0968444
performance	-.159709
management~r	-.1344091
p02	-.3385891
_Ip03_2	-1.86e-16
_Ip03_3	.1243056
p04	-.1732717
p05	.2395935
p06	-.5130172
p08	.1908796

An alternative way to interpret coefficients is to use structure function (Huberty, 1994) which measure the correlation between the discriminating variables and the discriminant functions, instead of standardized discriminant function coefficients. Again, p02, p06 and p05 are the most important discriminators, and behavior factor, p03 and p08 are not as important as them.

Since the question is not asking for complete discriminative analysis and only asks for the relevant discriminating variables, the question is answered and the rest is considered as extra work.

```
.
. estat classfunctions

Classification functions
```

	born_global	
	No	Yes
behaviour	-47.98996	-48.89752
performance	-22.05952	-21.56059
management~r	-37.51186	-37.14741
p02	-.4400063	-.3067625
_Ip03_2	-80.07746	-79.14614
_Ip03_3	-175.1338	-176.0024
p04	15.46637	14.78399
p05	26.35166	26.20161
p06	4.431713	4.49644
p08	.971671	.9599194
_cons	-25863.88	-25830.69
Priors	.5	.5

Alternatively, to Fisher approach to LDA, there is a Mahalanobis's approach to LDA using classification functions to calculate the probability of one observation belong to a group.

The complete observations, their true class and predicted class along with their probability of belonging to each group is shown in appendix III.

```
.
. estimates
```

---

active results

```
Linear discriminant analysis
Resubstitution classification summary
```

Key
Number Percent

True born_global	Classified No	Yes	Total
No	16	2	18
	88.89	11.11	100.00
Yes	2	12	14
	14.29	85.71	100.00
Total	18	14	32
	56.25	43.75	100.00
Priors	0.5000	0.5000	

Class table as a form of confusion matrix shows the posterior probabilities and absolute frequencies of the classifications and the true classes.

The accuracy for the discriminant function is 87.5%.

$$(Accuracy = \frac{True\ Positive + True\ Negative}{Positive + Negative})$$

II. Could you group the firms attending the three factors identified in case 1 and the % of exports in 2004? If you can, please, specify how many groups you would consider (and why) and characterize them.

First, we have to standardize our variables which we intend to use for clustering. The code is in appendix IV. **Then, since we don't know the number of groups, we have to choose the hierarchical method first and use Calinski method to obtain the number of groups.**

```
. cluster stop _clus_1, rule(calinski)
```

Number of clusters	Calinski/ Harabasz pseudo-F
2	6.52
3	6.19
4	9.00
5	8.76
6	12.05
7	12.27
8	12.69
9	12.20
10	12.95
11	12.99
12	13.18
13	13.05
14	13.06
15	13.45

Psuedo F describes the ratio of between cluster variance to within-cluster variance. If Psuedo F is increasing, that means either the within-cluster variance is decreasing (or constant while numerator increases) or enter-cluster variance is increasing(or constant while denominator decreases).

Here we have a swift increase in pseudo-F value suggesting number of groups to be 4, however, the value keeps increasing but the improvement is not consistent (there is a decrease in pseudo-F value after 4) and as a rule of thumb, we want to keep the groups as low as possible when the number of observations are low.

$$\rightarrow \boldsymbol{N.Groups = 4}$$

The cluster dendrogram is generated in appendix V to illustrate on the number of groups.

We can look at the simple statistics of the groups regarding the variables to see where there are possible distinctions:

```
. mean p06 behaviour performance management_behaviour, over(hierarg)

Mean estimation                                 Number of obs   =         49
```

	Mean	Std. Err.	[95% Conf.	Interval]
c.p06@hierarg				
1	16.5734	6.486541	3.531334	29.61547
2	37.35043	6.913448	23.45002	51.25085
3	66.75	12.99599	40.6198	92.8802
4	53.85714	13.41362	26.88725	80.82703
c.behaviour@hierarg				
1	-1.005895	.2197771	-1.447787	-.5640037
2	.2980793	.1386064	.0193926	.5767661
3	.3257197	.225824	-.1283298	.7797692
4	.9899608	.268062	.450986	1.528936
c.performance@hierarg				
1	-.4873786	.2374082	-.9647197	-.0100375
2	.6740244	.1206331	.4314752	.9165735
3	-1.193746	.439351	-2.077121	-.3103719
4	-.488128	.348597	-1.189029	.2127733
c.management_behaviour@hierarg				
1	.043036	.197327	-.3537165	.4397886
2	.1044284	.1968331	-.291331	.5001878
3	1.239051	.1910052	.8550094	1.623093
4	-1.143371	.3166219	-1.779982	-.5067599

As you can see, even though the averages are different, the confidence intervals have intersections asserting possible similarity in the population averages. So we run an ANOVA with Scheffe method for each group versus the other for each variable.

On the group characterization we can compare and test the variable means for each group:

```
. oneway performance hierarg

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      21.3801747      3   7.12672491     12.05     0.0000
 Within groups      26.6198262     45   .591551694
------------------------------------------------------------------------
    Total            48.000001     48   1.00000002

Bartlett's test for equal variances:  chi2(3) =   4.2689  Prob>chi2 = 0.234

. oneway performance hierarg, scheffe

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      21.3801747      3   7.12672491     12.05     0.0000
 Within groups      26.6198262     45   .591551694
------------------------------------------------------------------------
    Total            48.000001     48   1.00000002

Bartlett's test for equal variances:  chi2(3) =   4.2689  Prob>chi2 = 0.234

                Comparison of Scores for factor 2 by hierarg
                                (Scheffe)
Row Mean-
Col Mean |          1          2          3
---------+------------------------------------
       2 |     1.1614
         |      0.001
         |
       3 |   -.706368   -1.86777
         |      0.455      0.001
         |
       4 |   -.000749   -1.16215    .705618
         |      1.000      0.012      0.549
```

For example, ANOVA for behavior factor shows that the means for the groups are significantly different. To have more detail, with Scheffe method, we have significantly different means between group 1 versus 2, 2 versus 3, and 2 versus 4. The same comparison is shown for other factors and p06 in appendix VI.

Appendix

I. Code:

```
gen born_global = 1 if p09<=5
replace born_global = 0 if born_global==.
label variable born_global "Is the firm born global?"
label define born_global 1 "Yes" 0 "No"
label value born_global born_global
tabulate born_global
```

II. pooled within-group standardized, and total-sample standardized coefficients:

```
. estat loadings, all
```

Canonical discriminant function coefficients

	function1
behaviour	.4004514
performance	-.2201478
management~r	-.1608117
p02	-.0587926
_Ip03_2	-.410933
_Ip03_3	.3832848
p04	.3010959
p05	.0662079
p06	-.0285605
p08	.0051853
_cons	-14.78424

The standardized canonical discriminant function coefficients are the coefficients that apply to the discriminating variables after they have been standardized by the pooled within-group covariance.

Standardized canonical discriminant function coefficients

	function1
behaviour	.3797109
performance	-.1882737
management~r	-.1487195
p02	-1.065951
_Ip03_2	-.2122049
_Ip03_3	.1295739
p04	.1521327
p05	.680292
p06	-.8183387
p08	.2788845

III.

The complete observations, their true class and predicted class along with their probability of belonging to each group based on classification function.

```
. estat list
```

Obs.	Classification True	Class.	Probabilities No	Yes
1	Yes	Yes	0.2367	0.7633
2	Yes	Yes	0.0980	0.9020
3	Yes	Yes	0.0499	0.9501
4	Yes	Yes	0.0129	0.9871
7	Yes	Yes	0.0078	0.9922
8	Yes	Yes	0.0021	0.9979
9	No	No	0.9907	0.0093
10	No	No	0.9970	0.0030
12	No	No	0.9775	0.0225
13	No	Yes *	0.1107	0.8893
14	Yes	Yes	0.3260	0.6740
15	No	No	0.9657	0.0343
17	No	No	0.7145	0.2855
18	No	No	0.9579	0.0421
20	Yes	No *	0.5447	0.4553
21	Yes	Yes	0.0506	0.9494
22	No	No	0.5708	0.4292
24	No	No	0.7716	0.2284
25	No	No	0.9938	0.0062
27	Yes	Yes	0.4332	0.5668
29	No	No	0.8740	0.1260
30	No	Yes *	0.4249	0.5751
32	No	No	0.9002	0.0998
34	Yes	Yes	0.0242	0.9758
38	No	No	0.7761	0.2239
39	No	No	0.9978	0.0022
40	Yes	No *	0.5716	0.4284
41	No	No	0.7369	0.2631
42	Yes	Yes	0.0333	0.9667
43	Yes	Yes	0.0231	0.9769
47	No	No	0.9991	0.0009
48	No	No	0.7086	0.2914

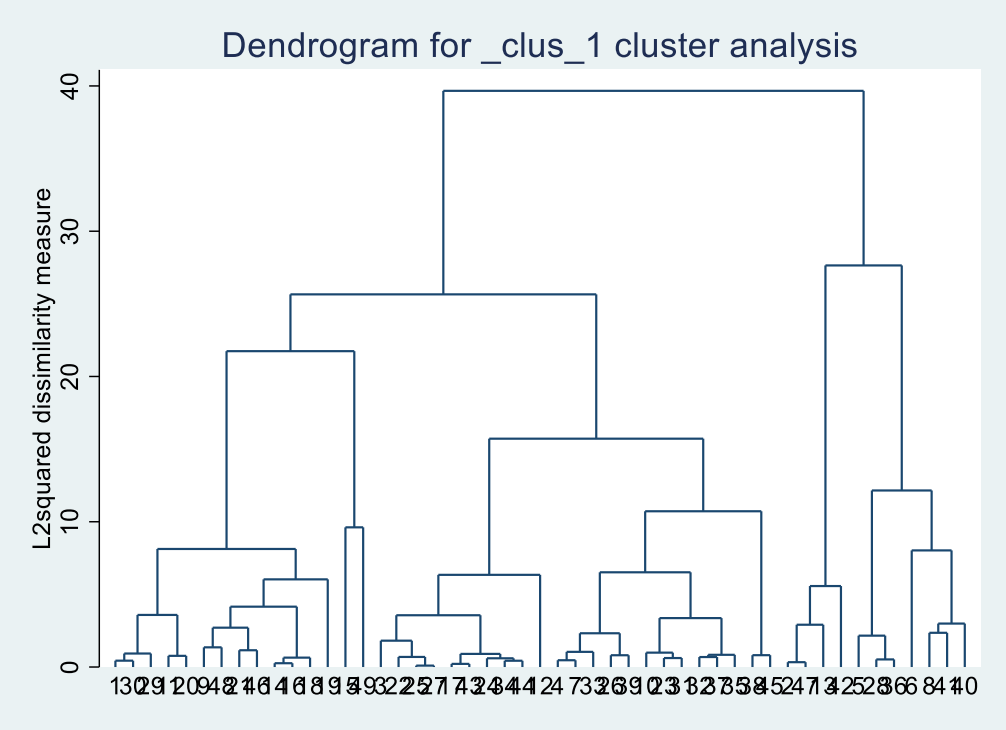
```
* indicates misclassified observations
```

IV. Code:

```
egen z2behaviour = std(behaviour)

egen z2performance = std(performance)

egen z2management_behaviour = std(management_behaviour)

egen z2p06  = std(p06)
```

cluster completelinkage z2behaviour z2performance z2management_behaviour z2p06 , measure(L2squared)

V. Dendrogram



VI. ANOVA for different groups

```
. oneway p06 hierarg, scheffe
```

Analysis of Variance

Source	SS	df	MS	F	Prob > F
Between groups	11714.122	3	3904.70734	4.12	0.0115
Within groups	42604.0624	45	946.756943		
Total	54318.1844	48	1131.62884		

Bartlett's test for equal variances: chi2(3) = 1.5861 Prob>chi2 = 0.663

Comparison of % of exports in 2004 by hierarg
(Scheffe)

Row Mean- Col Mean	1	2	3
2	20.777 0.261		
3	50.1766 0.051	29.3996 0.385	
4	37.2837 0.086	16.5067 0.674	-12.8929 0.930

Although the general null hypothesis of different mean for all groups are rejected, when we test each group with the other individually, almost none of the pairs have significant different p06 means.

```
. oneway performance hierarg, scheffe

                    Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      21.3801747       3   7.12672491     12.05     0.0000
 Within groups      26.6198262      45   .591551694
------------------------------------------------------------------------
    Total           48.000001      48   1.00000002

Bartlett's test for equal variances:  chi2(3) =   4.2689  Prob>chi2 = 0.234

                    Comparison of Scores for factor 2 by hierarg
                                    (Scheffe)
Row Mean-
Col Mean |          1          2          3
---------+---------------------------------
       2 |     1.1614
         |      0.001
         |
       3 |   -.706368   -1.86777
         |      0.455      0.001
         |
       4 |   -.000749   -1.16215    .705618
         |      1.000      0.012      0.549
```

The overall null hypothesis that the groups have different means regarding the variable performance is rejected. Comparing the means by each group and the other shows that group 1 versus 2, 2 versus 3, and 2 versus 4 have significantly different means.

```
. oneway management_behaviour hierarg, scheffe

                    Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      15.5706733       3   5.19022442     7.20     0.0005
 Within groups      32.4293275      45   .720651723
------------------------------------------------------------------------
    Total          48.0000008      48   1.00000002

Bartlett's test for equal variances:  chi2(3) =   3.0223  Prob>chi2 = 0.388

                    Comparison of Scores for factor 3 by hierarg
                                    (Scheffe)
Row Mean-
Col Mean |          1          2          3
---------+---------------------------------
       2 |    .061392
         |      0.997
         |
       3 |    1.19602    1.13462
         |      0.115      0.123
         |
       4 |   -1.18641    -1.2478   -2.38242
         |      0.036      0.015      0.001
```

overall null hypothesis that the groups have different means regarding the variable performance is rejected. Comparing the means by each group and the other shows that group 1 versus 4, 2 versus 4, and 3 versus 4 have significantly different means.This is particularly interesting since all the groups compared to the 4^{th} group have different means asserting that management_behaviour factor is effective at giving the fourth group a characteristic different than others helping the observations be distinguished in a separate group.

THE END