MAREB | Multivariate Analysis | Instructor: Josep Rialp

Students: Shayan Abbasi, Ivan Gutiérrez

---

1) FIRST QUESTION:

First things first, it seems that the dataset has some missing values. We can remove or impute the missing values, however, by doing so, the results have not changed. It seems that factor analysis package deals with the missing values internally. All that said, we can run the following commands to be sure.

```
. mdesc
```

Variable	Missing	Total	Percent Missing
p01	0	53	0.00
p02	1	53	1.89
p03	9	53	16.98
p04	6	53	11.32
p05	1	53	1.89
p06	1	53	1.89
p07_1	0	53	0.00
p07_2	0	53	0.00
p07_3	0	53	0.00
p07_4	0	53	0.00
p07_5	0	53	0.00
p07_6	0	53	0.00
p07_7	0	53	0.00
p07_8	1	53	1.89
p07_9	1	53	1.89
p07_10	2	53	3.77
p07_11	1	53	1.89
p07_12	1	53	1.89
p07_13	0	53	0.00
p08	13	53	24.53
p09	2	53	3.77

Missing value is a common problem in exploratory factor analysis; therefore, following Truxillo (2005) , Graham (2009), and Weaver and Maxwell (2014) proposed approach, we can remove the missing rows or try to impute them with the following commands:

```
. pwcorr p07_1-p07_13, sig star(0.05)
```

	p07_1	p07_2	p07_3	p07_4	p07_5	p07_6	p07_7
p07_1	1.0000						
p07_2	0.6884*	1.0000					
	0.0000						
p07_3	0.4850*	0.5292*	1.0000				
	0.0002	0.0000					
p07_4	0.4303*	0.5173*	0.6194*	1.0000			
	0.0013	0.0001	0.0000				
p07_5	0.3966*	0.5383*	0.5554*	0.6555*	1.0000		
	0.0033	0.0000	0.0000	0.0000			
p07_6	0.5107*	0.6424*	0.6377*	0.7112*	0.7205*	1.0000	
	0.0001	0.0000	0.0000	0.0000	0.0000		
p07_7	0.2826*	0.4131*	0.5381*	0.5283*	0.6175*	0.5607*	1.0000
	0.0403	0.0021	0.0000	0.0000	0.0000	0.0000	
p07_8	0.1957	0.2724	0.0779	0.0758	0.0788	0.1577	0.2131
	0.1644	0.0507	0.5832	0.5932	0.5789	0.2642	0.1293
p07_9	0.2457	0.4459*	0.1122	0.3216*	0.3873*	0.3064*	0.4794*
	0.0791	0.0009	0.4284	0.0201	0.0046	0.0272	0.0003
p07_10	0.5721*	0.5380*	0.4942*	0.4923*	0.5849*	0.5457*	0.7201*
	0.0000	0.0000	0.0002	0.0002	0.0000	0.0000	0.0000
p07_11	0.2856*	0.3610*	0.1048	0.3341*	0.2184	0.1681	0.4127*
	0.0401	0.0086	0.4595	0.0155	0.1198	0.2337	0.0024
p07_12	0.3778*	0.3680*	0.1216	0.4622*	0.4229*	0.3287*	0.3516*
	0.0058	0.0073	0.3905	0.0006	0.0018	0.0173	0.0106
p07_13	0.3163*	0.4195*	0.2383	0.5315*	0.6671*	0.4567*	0.2829*
	0.0210	0.0018	0.0857	0.0000	0.0000	0.0006	0.0401

	p07_8	p07_9	p07_10	p07_11	p07_12	p07_13
p07_8	1.0000					
p07_9	0.6012*	1.0000				
	0.0000					
p07_10	0.3706*	0.5329*	1.0000			
	0.0081	0.0001				
p07_11	0.3181*	0.5832*	0.5918*	1.0000		
	0.0229	0.0000	0.0000			
p07_12	0.2187	0.5384*	0.4082*	0.3880*	1.0000	
	0.1232	0.0000	0.0029	0.0049		
p07_13	-0.0594	0.2707	0.2253	0.1401	0.4900*	1.0000
	0.6758	0.0523	0.1120	0.3219	0.0002	

First, we can do a pairwise correlation which can show possible opportunity to reduce the dimension with factor extraction. We observe high and significant correlations between some of the variables confirming that. If we were going to use these variables in a regression model we could have multicollinearity, but that's not the problem for now. Here the main objective is to extract factors that reduce our dimension and help with the summarized interpretability.

Now, we want to check the feasibility of factor analysis using Kaiser Meyer Olkin test/Bartlett test. KMO test the hypothesis of whether the variables are correlated enough (and partially uncorrelated enough) for the factor analysis. The KMO statistic shows a good enough factor analysis (0.7<KMO<0.9). The higher the correlation and lower the partial correlation (correlation between two variables without the effect of other variables), the higher the KMO.

```
. factortest p07_1-p07_13

Determinant of the correlation matrix
Det                =     0.000


Bartlett test of sphericity

Chi-square         =           376.193
Degrees of freedom =                78
p-value            =             0.000
H0: variables are not intercorrelated


Kaiser-Meyer-Olkin Measure of Sampling Adequacy
KMO                =     0.834
```

Bartlett test, on the other hand, checks a certain redundancy between the variables that can be shown by factors. The rejected null hypothesis elaborates that the variables are not orthogonal (or correlated). In appendix I add an illustration of partial correlation, we can also see orthogonal intuition by looking at z vector.

KMO measures can also be calculated for each variable:

```
. estat kmo

Kaiser-Meyer-Olkin measure of sampling adequacy
```

Variable	kmo
p07_1	0.8401
p07_2	0.8599
p07_3	0.8548
p07_4	0.8658
p07_5	0.8503
p07_6	0.8897
p07_7	0.8659
p07_8	0.6756
p07_9	0.7738
p07_10	0.8404
p07_11	0.7780
p07_12	0.8438
p07_13	0.7338
Overall	0.8336

```
. factor p07_1-p07_13, pcf
(obs=49)

Factor analysis/correlation                        Number of obs    =         49
    Method: principal-component factors            Retained factors =          3
    Rotation: (unrotated)                          Number of params =         36
```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	6.12682	4.25263	0.4713	0.4713
Factor2	1.87419	0.73700	0.1442	0.6155
Factor3	1.13719	0.28734	0.0875	0.7029
Factor4	0.84985	0.10217	0.0654	0.7683
Factor5	0.74768	0.25237	0.0575	0.8258
Factor6	0.49530	0.00612	0.0381	0.8639
Factor7	0.48918	0.17822	0.0376	0.9016
Factor8	0.31096	0.05535	0.0239	0.9255
Factor9	0.25561	0.04023	0.0197	0.9451
Factor10	0.21538	0.02450	0.0166	0.9617
Factor11	0.19088	0.02286	0.0147	0.9764
Factor12	0.16802	0.02908	0.0129	0.9893
Factor13	0.13894	.	0.0107	1.0000

```
    LR test: independent vs. saturated:  chi2(78) =  384.98 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances
```

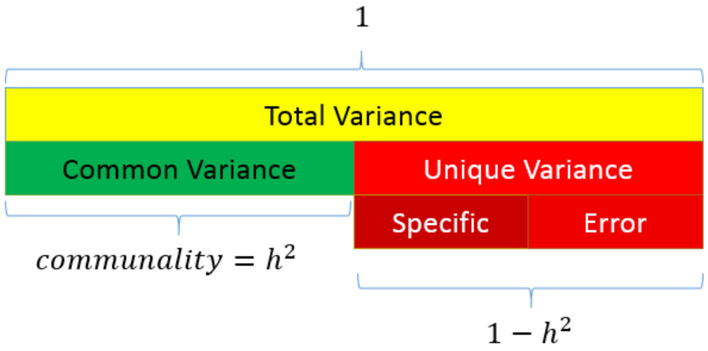
Variable	Factor1	Factor2	Factor3	Uniqueness
p07_1	0.7330	-0.0783	0.0031	0.4566
p07_2	0.7899	-0.0893	-0.0809	0.3615
p07_3	0.7133	-0.3361	-0.4129	0.2077
p07_4	0.7891	-0.2604	0.0501	0.3071
p07_5	0.8009	-0.3078	0.0966	0.2545
p07_6	0.7986	-0.3126	-0.1190	0.2503
p07_7	0.7362	0.0856	-0.3173	0.3500
p07_8	0.3341	0.7352	-0.0040	0.3478
p07_9	0.6071	0.6183	0.1876	0.2139
p07_10	0.7932	0.2652	-0.2638	0.2309
p07_11	0.4988	0.5612	-0.0481	0.4339
p07_12	0.5896	0.1747	0.5876	0.2766
p07_13	0.5541	-0.3755	0.6174	0.1708

In the next step we calculate the factor’s eigenvalues, which explain the total variability explained by each factor. Since we have 13 standard normalized variables, we have to explain variance of 13. So, if we divide the eigenvalue by the sum of eigenvalues, we get a proportion of variability explanation. The factors with eigenvalues more than 1 are usually be chosen, since they explain more than the variable itself. A geometric illustration of eigen values are given in appendix 2.

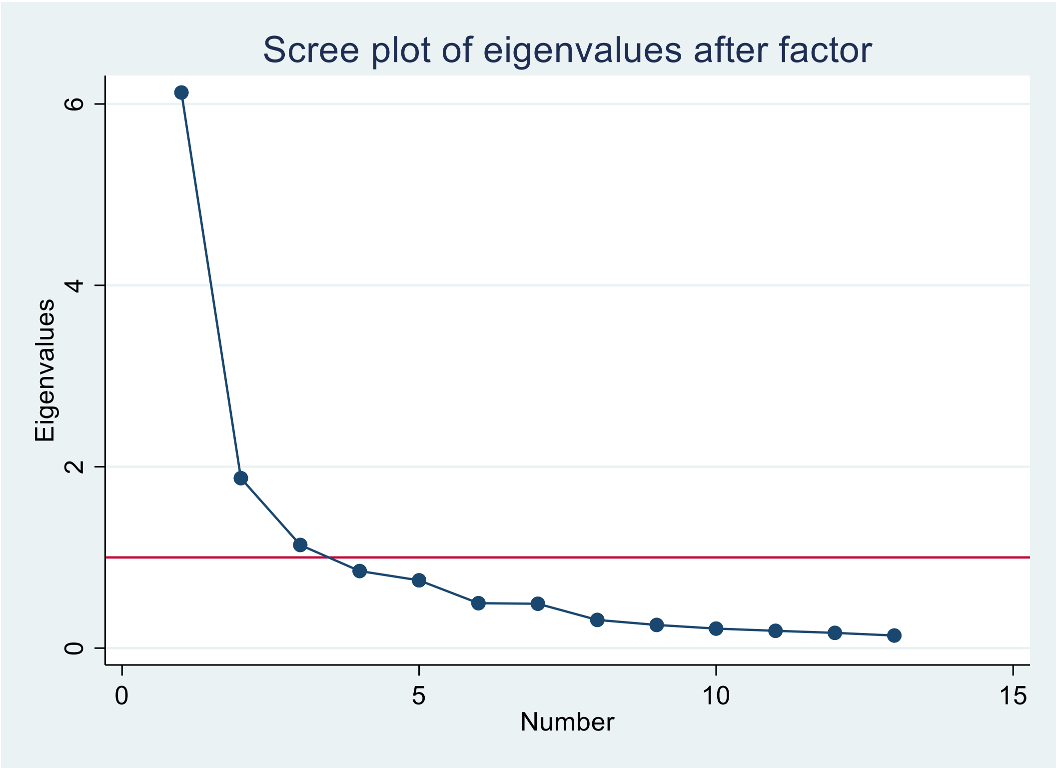
My interpretation of eigenvalue in geometric perspective is that eigenvalue is compatible with the mathematical interpretation of eigenvalue, which says that, if we transform a vector of the eigenvector line by the correlation matrix (or covariance matrix), the vector will change size (and not direction) by the eigenvalue amount.

The factor loadings above are the weights and correlations of variables with factors. High loading means that the variable is more related to the factor’s dimension. Negative load, on the contrary, shows an inverse impact on the factor. Moreover, the uniqueness shows how unique the variance is and not shared with other variables. Here, as an example, variable p07_1 shows 45.66% unique variance. High uniqueness shows lower importance of variable to the factor model and low uniqueness (or high communality) shows the relevance of the variable to factor model.

We can also show the Unique variance illustrated as: (From UCLA institute of digital research and education)



We can also show the corresponding eigenvalues of the factors by a plot:



```
. rotate, varimax horst blanks(.7)

Factor analysis/correlation                        Number of obs    =         49
    Method: principal-component factors            Retained factors =          3
    Rotation: orthogonal varimax (Kaiser on)       Number of params =         36
```

Factor	Variance	Difference	Proportion	Cumulative
Factor1	4.47492	1.83237	0.3442	0.3442
Factor2	2.64254	0.62181	0.2033	0.5475
Factor3	2.02074	.	0.1554	0.7029

```
    LR test: independent vs. saturated:  chi2(78) =  384.98 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances
```

Variable	Factor1	Factor2	Factor3	Uniqueness
p07_1				0.4566
p07_2	0.7067			0.3615
p07_3	0.8901			0.2077
p07_4	0.7085			0.3071
p07_5	0.7139			0.2545
p07_6	0.8138			0.2503
p07_7	0.7085			0.3500
p07_8		0.8073		0.3478
p07_9		0.8210		0.2139
p07_10				0.2309
p07_11		0.7191		0.4339
p07_12			0.7270	0.2766
p07_13			0.8551	0.1708

```
    (blanks represent abs(loading)<.7)

Factor rotation matrix
```

	Factor1	Factor2	Factor3
Factor1	0.8055	0.4251	0.4129
Factor2	-0.3692	0.9050	-0.2116
Factor3	-0.4636	0.0180	0.8859

When we want to make not inter-correlated components, we rotate the factor loads. Notice that now we just have 3 factors with a clearer view on results. The 3 factors explain 70.29% of the total variance observed.

Moreover, we can see each factor representing the underlying information on which variables. For example, Factor 2 is mostly defined by p07_8, p07_9 and p07_11.

Finally, we see the correlation matrix between the factors.

```
. predict behaviour performance management_behaviour
(option regression assumed; regression scoring)

Scoring coefficients (method = regression; based on varimax rotated factors)
```

Variable	Factor1	Factor2	Factor3
p07_1	0.11050	0.01311	0.06067
p07_2	0.15443	0.01042	0.00026
p07_3	0.32830	-0.11934	-0.23561
p07_4	0.13461	-0.07018	0.12158
p07_5	0.12654	-0.09151	0.16396
p07_6	0.21507	-0.09738	-0.00359
p07_7	0.20928	0.08738	-0.20721
p07_8	-0.09927	0.37812	-0.06361
p07_9	-0.11846	0.34363	0.11728
p07_10	0.15959	0.17890	-0.18198
p07_11	-0.02535	0.30482	-0.06720
p07_12	-0.19642	0.13457	0.47773
p07_13	-0.10486	-0.13308	0.56066

At the end, we can create the factors with the “predict” command. The table in the left side shows the coefficients of the regressions used to estimate them.

Interesting thing is, we can also mix the factors after each step. For example, imagine that if behaviour to management_behaviour has a meaning. Then we could have a new variable.

2) SECOND QUESTION:

a. Univariate Analysis

An ANOVA or a t-test can test the hypothesis of whether the average values of a factor (dimensions) are different between the family/non-family firms:

```
. oneway behaviour p04

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      .654602457      1   .654602457      0.69     0.4125
 Within groups       39.155718     41   .955017511
------------------------------------------------------------------------
    Total           39.8103204     42   .947864772

Bartlett's test for equal variances:  chi2(1) =   0.4262  Prob>chi2 = 0.514

. pwmean behaviour, over(p04) mcompare(tukey) effects

Pairwise comparisons of means with equal variances

over         : p04

note: option tukey ignored since there is only one comparison
```

behaviour	Contrast	Std. Err.	Unadjusted t	Unadjusted P>|t|	Unadjusted [95% Conf.	Interval]
p04						
Yes vs No	-.2468323	.2981391	-0.83	0.413	-.8489364	.3552718

Although the Bartlett test of equal variances is not rejected, we failed to reject the null hypothesis that the averages of the two categories are equal. Thus, we cannot make a conclusion with a 95% confidence interval. Pairwise mean comparison and the t-test shows the same thing.

```
. oneway  performance p04

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      3.78116441      1   3.78116441      3.99     0.0524
 Within groups      38.8445863     41   .947428934
------------------------------------------------------------------------
    Total           42.6257507     42   1.01489883

Bartlett's test for equal variances:  chi2(1) =   0.8785  Prob>chi2 = 0.349

. pwmean performance, over(p04) mcompare(tukey) effects

Pairwise comparisons of means with equal variances

over         : p04

note: option tukey ignored since there is only one comparison
```

performance	Contrast	Std. Err.	Unadjusted t	Unadjusted P>|t|	Unadjusted [95% Conf.	Interval]
p04						
Yes vs No	.593234	.2969522	2.00	0.052	-.0064732	1.192941

Although the Bartlett test of equal variances is not rejected, we failed to reject the null hypothesis that the averages of the two categories are equal. Thus, we cannot make a conclusion with a 95% confidence interval. Pairwise mean comparison and the t-test shows the same thing. However, with 90% confidence interval we can show that the average performance is different between the two categories.

```
. oneway management_behaviour p04

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      .049832779      1   .049832779      0.05     0.8201
 Within groups      39.0012347     41   .951249628
------------------------------------------------------------------------

    Total           39.0510675     42   .929787322

Bartlett's test for equal variances:  chi2(1) =   1.4879  Prob>chi2 = 0.223

. pwmean management_behaviour, over(p04) mcompare(tukey) effects

Pairwise comparisons of means with equal variances

over         : p04

note: option tukey ignored since there is only one comparison
```

management~r	Contrast	Std. Err.	Unadjusted t	Unadjusted P>|t|	Unadjusted [95% Conf.	Interval]
p04						
Yes vs No	.0681037	.2975504	0.23	0.820	-.5328114	.6690189

Although the Bartlett test of equal variances is not rejected, we failed to reject the null hypothesis that the averages of the two categories are equal. Thus, we cannot make a conclusion with a 95% confidence interval. Pairwise mean comparison and the t-test shows the same thing.

Extra Work: (not part of the assignment anymore)

With discriminate analysis we can derive a linear function that classifies the data (LDA is a supervised method), and in the process, the ANOVA is carried on too (which answers the second question). The idea in discriminant analysis is that maybe the different means in multivariate environment can be

First, we check the qualitative dependent variables imbalance:

```
. tabulate p04
```

Family firm	Freq.	Percent	Cum.
No	21	44.68	44.68
Yes	26	55.32	100.00
Total	47	100.00	

Then, we carry on the linear discriminant analysis and derive the confusion matrix:

```
. * discriminant analysis
. * priors set to default
. xi: discrim lda behaviour performance management_behaviour, group(p04) priors(0.5, 0.5)

Linear discriminant analysis
Resubstitution classification summary
```

Key
Number Percent

True p04	Classified No	Classified Yes	Total
No	13 61.90	8 38.10	21 100.00
Yes	7 31.82	15 68.18	22 100.00
Total	20 46.51	23 53.49	43 100.00
Priors	0.5000	0.5000	

```
. estat grsummarize
```

The mean of each factor for each class is shown.

Estimation sample discrim lda
Summarized by **p04**

Mean	p04 No	p04 Yes	Total
behaviour	.228161	-.0186713	.1018747
performance	-.3085855	.2846485	-.0050704
management~r	-.1304817	-.062378	-.095638
N	21	22	43

```
. estat correlations
```

The underlying factors are surely not correlated.

Pooled within-group correlation matrix

	behavi~r	perfor~e	manage~r
behaviour	1.00000		
performance	0.05198	1.00000	
management~r	0.03831	-0.04778	1.00000

The same results as we expected.

```
. estat anova

Univariate ANOVA summaries
```

Variable	Model MS	Resid MS	Total MS	R-sq	Adj. R-sq	F	Pr > F
behaviour	.65460246	39.155718	38.239025	0.0164	-0.0075	.68544	0.4125
performance	3.7811644	38.844586	38.009743	0.0887	0.0665	3.991	0.0524
managemen~r	.04983278	39.001235	38.07382	0.0013	-0.0231	.05239	0.8201

```
Number of obs = 43          Model df = 1        Residual df = 41
```

On the other hand, the coefficients of the linear function that classifies the data to family/non-family is the following:

```
. estat loadings, all

Canonical discriminant function coefficients
```

	function1
behaviour	.4346515
performance	-.9491559
management~r	-.1669674
_cons	-.065061

```
Standardized canonical discriminant function coefficients
```

	function1
behaviour	.4247632
performance	-.92387
management~r	-.1628467

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

	function1
behaviour	.4231695
performance	-.9562004
management~r	-.1609991

```
. estat structure

Canonical structure
```

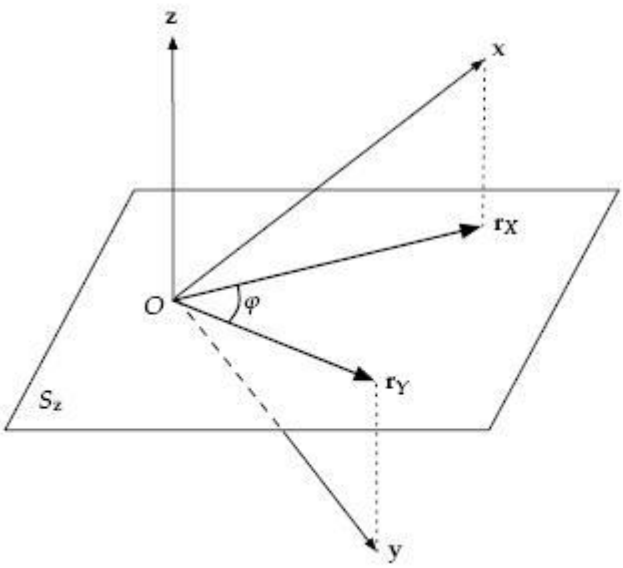
	function1
behaviour	.3704971
performance	-.8940074
management~r	-.1024264

```
. estat classfunctions

Classification functions

              |  p04
              |        No        Yes
--------------+----------------------
    behaviour |  .2631539  -.0331605
  performance | -.3472873   .2997795
 management~r | -.1638324  -.0500059
        _cons | -.0942932  -.0445351
--------------+----------------------
       Priors |        .5         .5
```

There is more intuition to Linear discriminant analysis, but for now, mentioning that ANOVA is carried on in the process of LDA is enough.

Appendix:

1. An illustration of partial correlation without the effect of variable z (vector z). Vector z is also orthogonal to the plane that contains the vectors showing the partial correlation. Therefore, rejecting orthogonal relationship with Bartlett test is another way to show correlation.



2. In factor analysis, the correlation matrix yields eigenvectors that are immune to change in angle by the transformation. The eigen value of the data now represents the variation of the projections of the data on the eigenvectors. An illustration is given in the following: