Exercises unit 1.3: Discriminant analysis

1) Previous analysis of the classification problem

Summary Statistics by Group

Group_3_classes	1	2	3	TOTAL
COUNTS	97	100	98	295
MEANS				
factorR	213,402	169,9	183,776	188,814
xcoord(east)	749548,	731143,	715616,	732037,
height	826,32	661,72	602,98	696,329
Coastdistance	28591,7	26480,9	25653,9	26900,3
STD. DEVIATIONS				
factorR	30,5807	25,9951	26,026	32,9546
xcoord(east)	17078,7	19423,9	13477,1	21760,9
height	322,958	364,159	285,36	338,329
Coastdistance	13758,0	15028,9	13181,6	14025,4

There are the most differences in xcoord.

In the first group there are the biggest values. In the second group there are the lowest values of factor and in the third group there are the lowest values of xcoord, height and Coastdistance.

Kovariance skupina 1 (group 1)

	xcoord(east)	factorR	height	Coastdistance
xcoord(east)	2,91682E8	76859,4	-4,07166E6	-2,21592E8
	(97)	(97)	(97)	(97)
factorR	76859,4	935,18	-1126,41	-110814,
	(97)	(97)	(97)	(97)
height	-4,07166E6	-1126,41	104302,	3,08501E6
	(97)	(97)	(97)	(97)
Coastdistance	-2,21592E8	-110814,	3,08501E6	1,89282E8
	(97)	(97)	(97)	(97)

If the means for a variable are significantly different in different groups, then we can say that this variable discriminates between the groups.

Kovariance skupina 2 (group 2)

	xcoord(east)	factorR	height	Coastdistance
xcoord(east)	3,77288E8	297770,	-5,40112E6	-2,76765E8
	(100)	(100)	(100)	(100)
factorR	297770,	675,747	-2086,29	-190370,
	(100)	(100)	(100)	(100)
height	-5,40112E6	-2086,29	132612,	4,83369E6
	(100)	(100)	(100)	(100)
Coastdistance	-2,76765E8	-190370,	4,83369E6	2,25868E8
	(100)	(100)	(100)	(100)

Kovariance skupina 3 (group 3)

	xcoord(east)	factorR	height	Coastdistance
xcoord(east)	1,81632E8	238567,	-3,14088E6	-1,74136E8
	(98)	(98)	(98)	(98)
factorR	238567,	677,351	-3145,8	-243280,
	(98)	(98)	(98)	(98)
height	-3,14088E6	-3145,8	81430,1	3,15848E6
	(98)	(98)	(98)	(98)
Coastdistance	-1,74136E8	-243280,	3,15848E6	1,73754E8
	(98)	(98)	(98)	(98)

Pooled Within-Group Statistics for Group_3_classes
Within-Group Covariance Matrix

	factorR	xcoord(east)	height	Coastdistance
factorR	761,573	205475,	-2122,67	-181791,
xcoord(east)	205475,	2,84148E8	-4,2132E6	-2,24533E8
height	-2122,67	-4,2132E6	106302,	3,70229E6
Coastdistance	-181791,	-2,24533E8	3,70229E6	1,96528E8

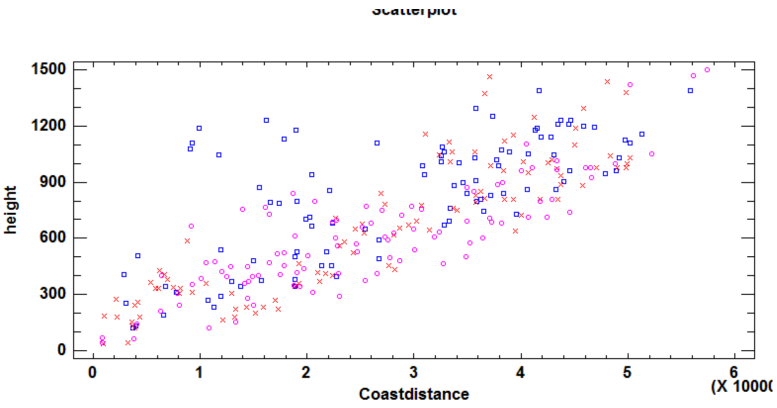
This pane shows the estimated correlations between the independent variables within each group. The within group information from all of the groups has been pooled.

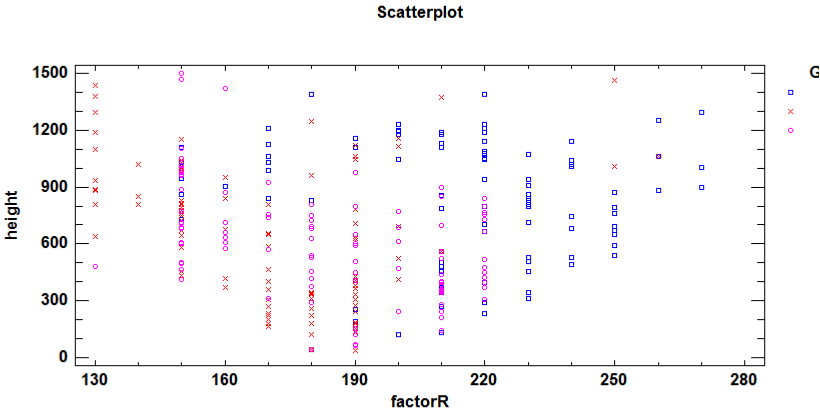
Within-Group Correlation Matrix

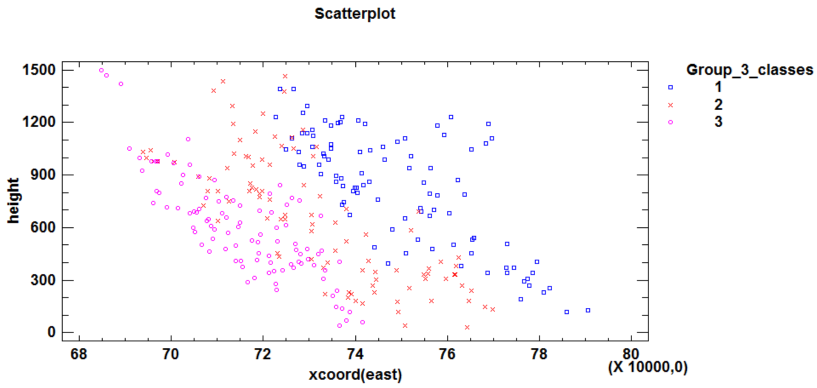
	factorR	xcoord(east)	height	Coastdistance
factorR	1,0	0,441704	-0,235915	-0,469898
xcoord(east)	0,441704	1,0	-0,7666	**-0,95016**
height	-0,235915	-0,7666	1,0	**0,810004**
Coastdistance	-0,469898	**-0,95016**	**0,810004**	1,0

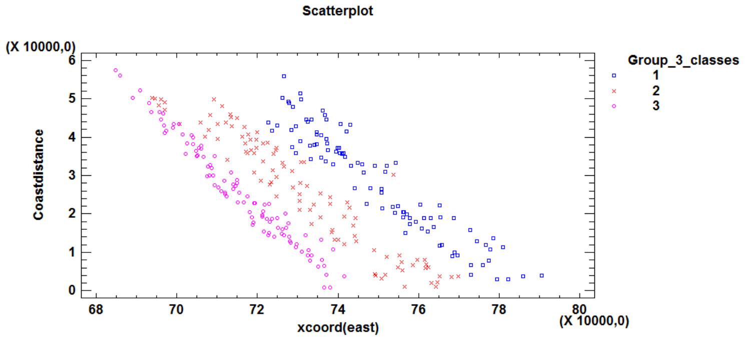
The Pooled Within-group Correlation matrix provides bivariate correlations between all variables. It can be used to detect potential problems with multicolliearity. It is necessary to pay attention if several correlation coefficient are larger than 0.8!

Usually, one includes several variables in a study in order to see which one(s) contribute to the discrimination between groups. In that case, we have a matrix of total variances and covariances -- we have a matrix of pooled within-group variances and covariances. We can compare those two matrices via multivariate *F* tests in order to determined whether or not there are any significant differences (with regard to all variables) between groups.


Scatterplot


Scatterplot


Scatterplot


Scatterplot

These variables have more discriminant power. FactorR dont separate, as well as height and coast distance dont separate groups. Between xcoord and coastDistance is the best combination. . There is quite good separation between xcoord and height. The rest of pairs of variables are not separated.

2) Fisher's Linear Discrimination Rule

Discriminant Function Coefficients for Group_2_classes

	1
xcoord(east)	1,97631
factorR	0,638688
height	-0,0549039
Coastdistance	2,12412

When the variables are in different units or have different variances, more insight is usually gained from the standardized coefficients. On the discriminant function more influence have variables with higher value of coefficient.

Unstandardized Coefficients

	1
xcoord(east)	0,000109811
factorR	0,022702
height	-0,000168222
Coastdistance	0,000151734
CONSTANT	-88,6365

Discriminant Score for the first 5 samples of data is following:

1,867

1,68351

1,86196

1,96055

2,02845

Group Centroids for Group_2_classes

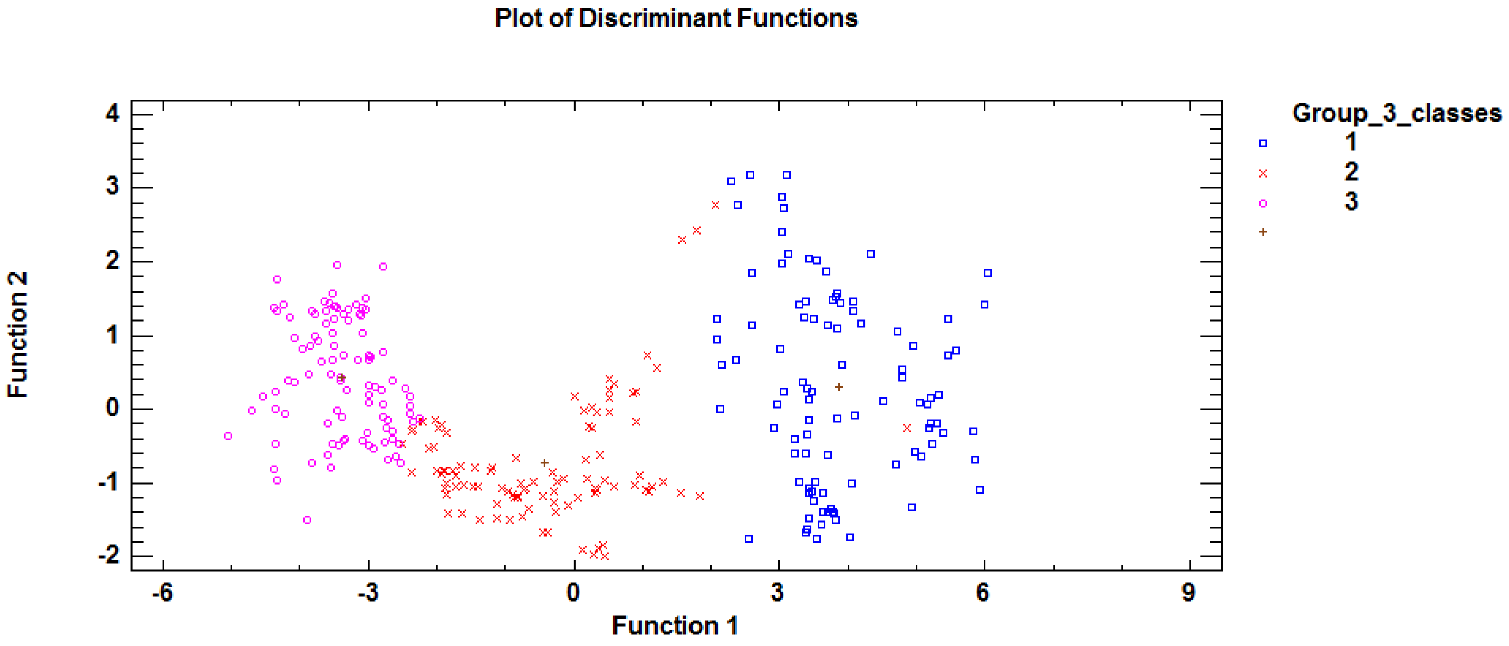
Group	1
1	2,71589
2	-1,33051

Group Centroids for Group_3_classes

Group	1	2
1	2,11175	0,326331
2	-0,16619	-0,727536
3	-1,92062	0,419383

Nº	DiscriminantFunctionValues 1	DiscriminantFunctionValues 2
1	-131,118	430,0182
2	-127,332	1293,794
3	-131,954	403,583
4	-127,434	1118,497
5	-129,529	760,3609

When I change a classification factor, I had to compute DiscriminantFunctionValue for two groups. This values are higher than in the previous.



3) Classification table

Classification Table

Actual	*Group*	*Predicted*	*Group_2_classes*
Group_2_classes	*Size*	*1*	*2*
1	97	97	0
		(100,00%)	(0,00%)
2	198	8	190
		(4,04%)	(95,96%)

Percent of cases correctly classified: 97,29%

	Prior
Group	*Probability*
1	0,5000
2	0,5000

This table shows the results of using the derived discriminant functions to classify observations. It lists the two highest scores amongst the classification functions for each of the 295 observations used to fit the model, as well as for any new observations. For example, row 1 scored highest for Group_2_classes = 1 and second highest for Group_2_classes = 2. In fact, the true value of Group_2_classes was 1. Amongst the 295 observations used to fit the model, 287 or 97,2881% were correctly classified. You can predict additional observations by adding new rows to the current data file, filling in values for each of the independent variables but leaving the cell for Group_2_classes blank.

	Actual	*Highest*	*Highest*	*Squared*		*2nd Highest*	*2nd Highest*	*Squared*	
Row	*Group*	*Group*	*Value*	*Distance*	*Prob.*	*Group*	*Value*	*Distance*	*Prob.*
1	1	1	4517,62	0,72061	0,9914	2	4512,86	10,2241	0,0086
2	1	1	4552,07	1,06579	0,9822	2	4548,06	9,08435	0,0178
3	1	1	4521,82	0,729191	0,9913	2	4517,09	10,1919	0,0087
4	1	1	4552,17	0,570533	0,9941	2	4547,04	10,8311	0,0059
5	1	1	4536,85	0,472566	0,9955	2	4531,45	11,2826	0,0045
6	1	1	4518,05	0,639385	0,9930	2	4513,1	10,5416	0,0070
7	1	1	4580,12	1,23535	0,9756	2	4576,43	8,61385	0,0244
8	1	1	4527,15	0,530543	0,9947	2	4521,91	11,0092	0,0053
9	1	1	4521,33	0,836028	0,9889	2	4516,84	9,80975	0,0111
10	1	1	4532,03	0,328053	0,9972	2	4526,16	12,0662	0,0028
11	1	1	4563,69	0,796239	0,9898	2	4559,11	9,94819	0,0102

Classification Function Coefficients for Group_2_classes

	1	*2*
xcoord(east)	0,011461	0,0110167
factorR	1,24801	1,15614
height	-0,0756357	-0,074955
Coastdistance	0,0156989	0,0150849
CONSTANT	-4622,34	-4260,87

Prior Probabilities: method for determining the probability of group membership before the data is examined.

Classification Table

Actual	*Group*	*Predicted*	*Group_2_classes*
Group_2_classes	*Size*	*1*	*2*
1	97	97	0
		(100,00%)	(0,00%)
2	198	6	192
		(3,03%)	(96,97%)

Percent of cases correctly classified: 97,97%

	Prior
Group	*Probability*
1	0,3288
2	0,6712

When we change Prior Probabilities to "Proportional to observed" we can see that Prior Probability in each group change from 0,5 for first group and 0,5 for second group to 0,3288 for the first group and 0,6712 to a second group. The probability of belonging to each group can help us to review.

	Actual	*Highest*	*Highest*	*Squared*		*2nd Highest*	*2nd Highest*	*Squared*	
Row	*Group*	*Group*	*Value*	*Distance*	*Prob.*	*Group*	*Value*	*Distance*	*Prob.*
1	1	1	4517,2	0,72061	0,9827	2	4513,16	10,2241	0,0173
2	1	1	4551,65	1,06579	0,9643	2	4548,35	9,08435	0,0357
3	1	1	4521,4	0,729191	0,9823	2	4517,39	10,1919	0,0177
4	1	1	4551,75	0,570533	0,9881	2	4547,33	10,8311	0,0119
5	1	1	4536,43	0,472566	0,9909	2	4531,74	11,2826	0,0091
6	1	1	4517,64	0,639385	0,9858	2	4513,4	10,5416	0,0142

Classification Table

Actual	*Group*	*Predicted*	*Group_2_classes*
Group_2_classes	*Size*	*1*	*2*
1	97	97	0
		(100,00%)	(0,00%)
2	198	6	192
		(3,03%)	(96,97%)

Percent of cases correctly classified: 97,97%

	Actual	*Highest*	*Highest*	*Squared*		*2nd Highest*	*2nd Highest*	*Squared*	
Row	*Group*	*Group*	*Value*	*Distance*	*Prob.*	*Group*	*Value*	*Distance*	*Prob.*
99	2	*1	4533,16	0,0392754	0,9987	2	4526,48	14,8088	0,0013
100	2	*1	4688,04	0,184151	0,9999	2	4678,84	20,0303	0,0001
101	2	*1	4513,47	0,275725	0,9953	2	4508,12	12,3996	0,0047
102	2	*1	4468,32	0,48864	0,9905	2	4463,68	11,2049	0,0095
109	2	*1	4413,56	2,90061	0,6414	2	4412,98	5,49097	0,3586
116	2	*1	4465,72	2,80696	0,6668	2	4465,03	5,62164	0,3332

* = incorrectly classified.

Classification Table

Actual	*Group*	*Predicted*	*Group_3_classes*	
Group_3_classes	*Size*	*1*	*2*	*3*
1	97	97	0	0
		(100,00%)	(0,00%)	(0,00%)
2	100	4	87	9
		(4,00%)	(87,00%)	(9,00%)
3	98	0	0	98
		(0,00%)	(0,00%)	(100,00%)

Percent of cases correctly classified: 95,59%

	Actual	*Highest*	*Highest*	*Squared*		*2nd Highest*	*2nd Highest*	*Squared*	
Row	*Group*	*Group*	*Value*	*Distance*	*Prob.*	*Group*	*Value*	*Distance*	*Prob.*
99	2	*1	10843,8	9,28612	0,9891	2	10839,3	18,3548	0,0109
100	2	*1	11336,4	1,30928	1,0000	2	11323,1	28,0844	0,0000
101	2	*1	10812,4	8,78308	0,9509	2	10809,4	14,7727	0,0491
102	2	*1	10754,9	9,15478	0,8839	2	10752,9	13,2765	0,1160
181	2	*3	10334,7	2,51325	0,5065	2	10334,7	2,60529	0,4935
182	2	*3	10328,0	2,61737	0,5256	2	10327,9	2,8626	0,4744
188	2	*3	10286,8	2,73539	0,6210	2	10286,3	3,76338	0,3790
191	2	*3	10300,3	1,80245	0,6912	2	10299,5	3,45395	0,3088
192	2	*3	10341,3	2,20147	0,5831	2	10341,0	2,91297	0,4169
193	2	*3	10275,3	1,59105	0,7579	2	10274,2	3,91434	0,2421
194	2	*3	10286,2	1,60435	0,7425	2	10285,1	3,76319	0,2575
196	2	*3	10274,7	1,56383	0,7685	2	10273,5	4,00379	0,2315
197	2	*3	10254,7	1,61196	0,7984	2	10253,4	4,40528	0,2016

* = incorrectly classified.

New datum:::

	Actual	*Highest*	*Highest*	*Squared*		*2nd Highest*	*2nd Highest*	*Squared*	
Row	*Group*	*Group*	*Value*	*Distance*	*Prob.*	*Group*	*Value*	*Distance*	*Prob.*
296		3	9200,83	39,3206	1,0000	2	9176,93	87,1099	0,0000

4) Discriminant power of the variables

Discriminant	*Eigenvalue*	*Relative*	*Canonical*
Function		*Percentage*	*Correlation*
1	8,87689	96,95	0,94803
2	0,278811	3,05	0,46693

The 2 discriminant functions take P-values less than 0.05 and are therefore statistically significant at the 95,0% confidence level.

Functions	*Wilks*			
Derived	*Lambda*	*Chi-Square*	*DF*	*P-Value*
1	0,0791723	736,7453	8	0,0000
2	0,781977	71,4428	3	0,0000

Values of Canonical Correlation and Wilks' Lambda indicate that first discriminant function has a greater discriminanting power because the calue of Wilks Lambda is not close to 1 and Canonical Correlation is close to 1.

5) Selection of variables

```
Stepwise regression
Method: forward selection
F-to-enter: 4,0
F-to-remove: 4,0

    Step 0:
    0 variables in the model.

    Step 1:
    Adding variable xcoord(east) with F-to-enter = 98,9773
    1 variables in the model.
    Wilk's lambda = 0,595974  Approximate F = 98,9773 with P-value =
0,0000

    Step 2:
    Adding variable Coastdistance with F-to-enter = 673,957
    2 variables in the model.
    Wilk's lambda = 0,105819  Approximate F = 301,782 with P-value =
0,0000

    Step 3:
    Adding variable factorR with F-to-enter = 47,4061
    3 variables in the model.
    Wilk's lambda = 0,0797467  Approximate F = 245,644 with P-value =
0,0000

    Final model selected.
```

```
Stepwise regression
Method: backward selection
F-to-enter: 4,0
F-to-remove: 4,0

    Step 0:
    4 variables in the model.
    Wilk's lambda = 0,0791723  Approximate F = 184,524 with P-value
= 0,0000

    Step 1:
    Removing variable height with F-to-remove = 1,04832
    3 variables in the model.
    Wilk's lambda = 0,0797467  Approximate F = 245,644 with P-value
= 0,0000

    Final model selected.
```

Forward selection:

Classification Table

Actual	*Group*	*Predicted*	*Group_3_classes*	
Group_3_classes	*Size*	*1*	*2*	*3*
1	97	97	0	0
		(100,00%)	(0,00%)	(0,00%)
2	100	4	89	7
		(4,00%)	(89,00%)	(7,00%)
3	98	0	0	98
		(0,00%)	(0,00%)	(100,00%)

Percent of cases correctly classified: 96,27%

All variables:

Classification Table

Actual	*Group*	*Predicted*	*Group_3_classes*	
Group_3_classes	*Size*	*1*	*2*	*3*
1	97	97	0	0
		(100,00%)	(0,00%)	(0,00%)
2	100	4	85	11
		(4,00%)	(85,00%)	(11,00%)
3	98	0	0	98
		(0,00%)	(0,00%)	(100,00%)

Percent of cases correctly classified: 94,92%

.

In order to compare processes in making discriminant analysis it is enough to look at percent of cases correctly classified. In Forward selection, there is 96,3 % correctly classified, in all variables classification process it is 95 %.