Fifth Homework- Data Analysis Course of Professor Murthi on Clustering, Discriminant Analysis, Simultaneous equations, and PCA

Meisam Hejazinia

03/05/2013

1 First Question: Simultaneous equation

Part a

Both 2SLS model and OLS was run over data, and the result is shown in the following table. While 2SLS estimated 7.54 as a standardized coefficient, OLS is estimating 12.60, showing that if we use OLS we will have upward bias. The bis is $E(\hat{\beta}) - \beta = 12.60 - 7.54 = 5.06$.

Part b

3SLS was run on the model, and as could be seen in the table the estimates of 3SLS are different from 2SLS. This shows even 2SLS was biased due to SUR mainly between the first two equations error term correlation. The bias of 2SLS in comparison to 3SLS is $7.54 - 6.20 = 1.34$. The correlation table given below shows error terms of market share and relative quality are highly correlated. The product line width error term is highly correlated with error term ofrelative quality. The rest two indogeneous variable's error terms, mean price, and relative direct cost are also correlated but with lower ratio, with themsleves and the other indogeneous variables error term.

3SLS estimate	2SLS estimate	OLS estimate
6.20	7.54	12.60

Table 1: Effect of Pioneering on market share

Variable	2SLS	3SLS
Intercept	42.00	-9.79
QUAL	0.51	0.47
PLB	-1.01	-0.90
PRICE	0.85	1.27
PION	7.54	6.20
TYRP	-0.38	0.98
EF	5.79	5.85
PHPF	0.58	-0.51
PLPF	0.17	1.92
PSC	-30.90	-30.32
PAPC	-1.51	-0.12
NCOMP	-7.54	-8.25
MKTEXP	-0.29	-0.33

Equations	MS	QUAL	PLB	PRICE	DC
MS	315.01	-193.20	24.13	9.83	9.51
QUAL	-193.20	601.57	-63.37	-105.72	-19.65
PLB	24.13	-63.37	79.73	16.74	2.80
PRICE	9.83	-105.72	16.74	32.11	1.31
DC	9.51	-19.65	2.80	1.31	1.12

Table 2: 3SLS correlation between error terms of indogeneous variables

2 Second question: Factor analysis

As could be seen in the following table which is created using 'proc corr' in SAS most of the variables are correlated. Q17, and Q18 are almost correlated with all Q1-Q19 except for Q15, Q8, Q9 and Q12. Q17 is also not correlated with Q1,Q3,Q4,Q5,Q6,Q7,Q8,Q9. Q16 is correlated with Q1,Q2,Q7, and Q16. Q19 is correlated with Q2 and Q9. Q15 is correlated with Q7. There are other correlated questions which are shown in the table with p-value lower than .05.

Part a To select the number of factors to retain we will look at the deminishing return. Probably two factors would be enough, since together they capture around 57% of the variation. On the other hand we can use measure of $1/20 = .05$ for selection, since here we have around 20 variables. This measure tells us that 5 factors are enough since the proportion of eigen vector greater than .05 condition will jeopardize on sixth eigen value.

Part b According to scree plot, if we select 2 factors due to diminishing return we will acount for 57% of the variation, yet if we select 5 factors based on the measure of number of variables we will have 75% of the variation captured by five factors.

Par c The first factor could be called reliability, since according to the table it is heavily loaded by intelligence, likable, expertise, competence, knowledge, and interesting attributes. The second factor could be called charm since it is heavily loaded by attractiveness, and beauty. Second could be called conviviality, since it is heavily loaded by friendliness, and confidence. The fourth factor could be named service quality, since expertise, and friendliness play some how role with same weight in it. Finally the last factor could be named exciting since activeness, exciting, and likable play significant role in it.

Part d: Based on commonality Q4 is not well represented, since only .55% of its variation is explained by five factors. The rest have more than .70% explanation by five factors.

Part e: Result of varimax rotation shows that, first factor is heavily loaded on Q17,Q10,Q19, Q13, Q16, Q18, and negatively on Q4, Q1, Q2, Q11, and Q14. Means it is positively heavily loaded on trustworthy, believability, being interesting, identify with, not irritating, and sincereness, likeable, knowledgable, intelligent. As we search these on google we see the word 'Reliablity', so probably the first factor could be selected as reliability. Second factor is positively highly loaded by Q3, and Q12, mean attractiveness, and beauty, and negatively loaded highly by Q6, mean good looking. All these could be summerized into the word charm. Third factor is hevaily positively loaded into Q7, and Q5, mean similar to you, exciting and heavily negatively loaded into Q15, mean active. As a result factor 3 could be labeled into conviviality. Factor 4 is highly negatively loaded into Q11, and Q9, mean friendly, expertise. I will call them service aspect, since on the service delivery you need to be professional both in human and in functional aspect. Finally fifth factor is heavily loaded on Q9, and Q14, mean unfriendliness, not competent. This could be labeld as disqualification (Disq).
Again regarding the communality, explanation of variance, intelligence is less explained by the factors relative to others, and trustworthiness is highly explained by them.
The number of factors to retain would not change, since vermiax rotation only tries to disperse the variances. The percentage of the variance explanation would also be the same.

3 Question 3: PDA Clustering and Discriminant Analysis

First part Proc cluster gives us the hierarchical clustering. The problem is that number of clusters (NCL) is 159. Only homogeneous clusters could be coupled together, mean R^2 close to one. Due to the size of the table I did not put it in the neat way

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Q1	1.00	0.55	-0.40	0.43	-0.32	0.42	-0.38	0.04	0.24	-0.42
	¡.0001	0.00	0.00	0.02	0.00	0.01		0.78	0.09	0.00
Q2	0.55	1.00	-0.28	0.68	-0.36	0.55	-0.49	0.42	0.32	-0.46
	¡.0001	0.05	¡.0001	0.01	¡.0001	0.00		0.00	0.03	0.00
Q3	-0.40	-0.28	1.00	-0.17	0.43	-0.69	0.34	-0.03	-0.05	-0.02
	0.00	0.05	0.23	0.00	¡.0001	0.02		0.83	0.74	0.89
Q4	0.43	0.68	-0.17	1.00	-0.24	0.39	-0.27	0.49	0.26	-0.45
	0.00	¡.0001	0.23	0.10	0.01	0.06		0.00	0.07	0.00
Q5	-0.32	-0.36	0.43	-0.24	1.00	-0.40	0.42	0.08	-0.02	0.29
	0.02	0.01	0.00	0.10	0.00	0.00		0.60	0.92	0.04
Q6	0.42	0.55	-0.69	0.39	-0.40	1.00	-0.31	0.29	0.17	-0.12
	0.00	¡.0001	¡.0001	0.01	0.00	0.03		0.04	0.23	0.42
Q7	-0.38	-0.49	0.34	-0.27	0.42	-0.31	1.00	-0.24	-0.33	0.40
	0.01	0.00	0.02	0.06	0.00	0.03		0.09	0.02	0.00
Q8	0.04	0.42	-0.03	0.49	0.08	0.29	-0.24	1.00	0.31	-0.23
	0.78	0.00	0.83	0.00	0.60	0.04	0.09	0.03	0.11	0.00
Q9	0.24	0.32	-0.05	0.26	-0.02	0.17	-0.33	0.31	1.00	-0.33
	0.09	0.03	0.74	0.07	0.92	0.23	0.02	0.03	0.02	0.21
Q10	-0.42	-0.46	-0.02	-0.45	0.29	-0.12	0.40	-0.23	-0.33	1.00
	0.00	0.00	0.89	0.00	0.04	0.42	0.00	0.11	0.02	0.00
Q11	0.29	0.61	-0.13	0.39	-0.16	0.46	-0.53	0.45	0.18	-0.41
	0.04	¡.0001	0.36	0.01	0.26	0.00	¡.0001	0.00	0.21	0.00
Q12	-0.41	-0.43	0.66	-0.21	0.38	-0.74	0.33	-0.13	-0.01	-0.01
	0.00	0.00	¡.0001	0.14	0.01	¡.0001	0.02	0.35	0.97	0.94
Q13	-0.55	-0.46	0.37	-0.43	0.55	-0.48	0.45	-0.15	-0.22	0.55
	¡.0001	0.00	0.01	0.00	¡.0001	0.00	0.00	0.31	0.12	¡.0001
Q14	0.50	0.61	-0.30	0.52	-0.12	0.37	-0.35	0.33	0.52	-0.39
	0.00	¡.0001	0.04	0.00	0.43	0.01	0.01	0.02	¡.0001	0.01
Q15	0.32	0.40	-0.27	0.28	-0.41	0.33	-0.56	0.19	0.29	-0.21
	0.02	0.00	0.05	0.05	0.00	0.02	¡.0001	0.18	0.04	0.14
Q16	-0.60	-0.58	0.35	-0.43	0.30	-0.40	0.52	-0.30	-0.35	0.44
	¡.0001	¡.0001	0.01	0.00	0.03	0.00	¡.0001	0.04	0.01	0.00
Q17	-0.43	-0.68	-0.02	-0.58	0.14	-0.29	0.42	-0.38	-0.38	0.68
	0.00	¡.0001	0.88	¡.0001	0.32	0.04	0.00	0.01	0.01	¡.0001
Q18	-0.59	-0.69	0.47	-0.52	0.51	-0.53	0.63	-0.30	-0.39	0.45
	¡.0001	¡.0001	0.00	¡.0001	0.00	¡.0001	¡.0001	0.03	0.00	0.00
Q19	-0.40	-0.60	-0.10	-0.38	0.16	-0.13	0.35	-0.31	-0.54	0.72
	0.00	¡.0001	0.49	0.01	0.26	0.38	0.01	0.03	¡.0001	¡.0001

Table 3: Survey question correlations

	Q11	Q12	Q13	Q14	Q16	Q17	Q18	Q19	
Q1	0.29	-0.41	-0.55	0.50	0.32	-0.60	-0.43	-0.59	-0.40
	0.04	0.00	¡.0001	0.00	¡.0001	0.00	¡.0001	0.00	
Q2	0.61	-0.43	-0.46	0.61	0.40	-0.58	-0.68	-0.69	-0.60
	¡.0001	0.00	0.00	¡.0001	¡.0001	¡.0001	¡.0001	¡.0001	
Q3	-0.13	0.66	0.37	-0.30	-0.27	0.35	-0.02	0.47	-0.10
	0.36	¡.0001	0.01	0.04	0.01	0.88	0.00	0.49	
Q4	0.39	-0.21	-0.43	0.52	0.28	-0.43	-0.58	-0.52	-0.38
	0.01	0.14	0.00	0.00	0.00	¡.0001	¡.0001	0.01	
Q5	-0.16	0.38	0.55	-0.12	-0.41	0.30	0.14	0.51	0.16
	0.26	0.01	¡.0001	0.43	0.03	0.32	0.00	0.26	
Q6	0.46	-0.74	-0.48	0.37	0.33	-0.40	-0.29	-0.53	-0.13
	0.00	¡.0001	0.00	0.01	0.00	0.04	¡.0001	0.38	
Q7	-0.53	0.33	0.45	-0.35	-0.56	0.52	0.42	0.63	0.35
	¡.0001	0.02	0.00	0.01	¡.0001	0.00	¡.0001	0.01	
Q8	0.45	-0.13	-0.15	0.33	0.19	-0.30	-0.38	-0.30	-0.31
	0.35	0.31	0.02		0.04	0.01	0.03	0.03	
Q9	0.18	-0.01	-0.22	0.52	0.29	-0.35	-0.38	-0.39	-0.54
	0.97	0.12	¡.0001		0.01	0.01	0.00	¡.0001	
Q10	-0.41	-0.01	0.55	-0.39	-0.21	0.44	0.68	0.45	0.72
	0.94	¡.0001	0.01		0.00	¡.0001	0.00	¡.0001	
Q11	1.00	-0.40	-0.51	0.42	0.25	-0.50	-0.69	-0.56	-0.49
	0.00	0.00	0.00		0.00	¡.0001	¡.0001	0.00	
Q12	-0.40	1.00	0.38	-0.22	-0.23	0.39	0.11	0.50	0.05
	0.00	0.01	0.12		0.01	0.43	0.00	0.76	
Q13	-0.51	0.38	1.00	-0.48	-0.50	0.54	0.58	0.63	0.46
	0.00	0.01	0.00		¡.0001	¡.0001	¡.0001	0.00	
Q14	0.42	-0.22	-0.48	1.00	0.32	-0.67	-0.59	-0.62	-0.58
	0.00	0.12	0.00		¡.0001	¡.0001	¡.0001	¡.0001	
Q15	0.25	-0.23	-0.50	0.32	1.00	-0.33	-0.30	-0.45	-0.28
	0.08	0.10	0.00	0.02	0.03	0.00	0.05		
Q16	-0.50	0.39	0.54	-0.67	-0.33	1.00	0.63	0.74	0.51
	0.00	0.01	¡.0001	¡.0001	¡.0001	¡.0001	0.00		
Q17	-0.69	0.11	0.58	-0.59	-0.30	0.63	1.00	0.60	0.76
	¡.0001	0.43	¡.0001	¡.0001	¡.0001	¡.0001	¡.0001		
Q18	-0.56	0.50	0.63	-0.62	-0.45	0.74	0.60	1.00	0.52
	¡.0001	0.00	¡.0001	¡.0001	¡.0001	¡.0001	0.00		
Q19	-0.49	0.05	0.46	-0.58	-0.28	0.51	0.76	0.52	1.00
	0.00	0.76	0.00	¡.0001	0.00	¡.0001	0.00		

Table 4: Survey question correlations

	Eigenvalue	Difference	Proportion	Cumulative
1	8.273	5.715	0.435	0.435
2	2.558	1.178	0.135	0.570
3	1.380	0.328	0.073	0.643
4	1.051	0.030	0.055	0.698
5	1.022	0.205	0.054	0.752
6	0.817	0.206	0.043	0.795
7	0.611	0.065	0.032	0.827
8	0.546	0.049	0.029	0.856
9	0.496	0.046	0.026	0.882
10	0.450	0.086	0.024	0.905
11	0.364	0.044	0.019	0.925
12	0.320	0.074	0.017	0.941
13	0.246	0.016	0.013	0.954
14	0.231	0.035	0.012	0.967
15	0.196	0.059	0.010	0.977
16	0.137	0.022	0.007	0.984
17	0.115	0.011	0.006	0.990
18	0.104	0.019	0.006	0.996
19	0.085		0.005	1.000

Table 5: Principle Component Analysis

	Reliable	Charm	Convivial	Service quality	Exciting
¬ dull	0.873	0.12549	0.05974	-0.06811	0.01055
¬ irritating	0.79015	-0.02104	0.01504	-0.13888	0.17411
¬ Not trustworthy	0.77794	-0.45606	-0.00609	0.21173	0.07989
¬ Dont identify with	0.75763	0.12186	0.31435	0.16802	0.02869
¬ not sincere	0.68057	-0.547	0.15761	-0.02032	0.08602
¬ Unexciting	0.66838	0.09973	0.20596	-0.03892	-0.46597
¬ not believable	0.62429	-0.43632	0.33732	0.22826	0.10345
¬ Not similar to you	0.48119	0.45229	0.47944	0.19992	-0.13484
¬ good looking	-0.63025	-0.53581	0.3323	-0.0385	-0.03315
¬ Intelligent	-0.66397	0.1566	0.25673	-0.12906	-0.09253
¬ Likeable	-0.67973	-0.15167	-0.19626	0.13379	-0.41977
¬ Expert	-0.69626	0.11874	0.25657	-0.39877	0.15329
¬ competent	-0.73442	0.193	0.13346	0.37164	-0.22254
¬ Knowledgeable	-0.83463	0.05775	0.18118	-0.11549	-0.04003
¬ Unattractive	0.43777	0.74726	-0.08553	-0.21189	0.09847
¬ Ugly	0.50899	0.6647	-0.25446	0.08194	0.08883
¬ confident	-0.43317	0.2725	0.62693	-0.0436	0.35427
¬ friendly	-0.46896	0.35555	0.02933	0.65096	0.14664
¬ active	-0.54508	-0.15601	-0.26226	0.19145	0.55284

Table 6: Five loaded factor labeled

	Reliability	Charm	conviviality	service aspect	Disq
Trustworthy	0.84	0.01	0.13	-0.35	-0.15
Believable	0.83	-0.13	0.23	-0.03	-0.04
Sincere	0.78	-0.13	0.16	-0.16	-0.34
Identify with	0.62	0.36	0.45	0.03	0.02
¬ irritable	0.59	0.41	0.18	-0.10	-0.35
Interesting	0.55	0.49	0.39	-0.15	-0.26
¬ intelligent	-0.54	-0.27	-0.02	0.41	0.14
¬ likeable	-0.58	-0.51	-0.09	-0.22	0.24
¬ knowledgable	-0.61	-0.39	-0.18	0.40	0.16
Attractive	-0.07	0.86	0.22	0.09	-0.11
Beautiful	0.06	0.85	0.15	-0.16	0.09
Bad Looking	-0.14	-0.81	-0.17	0.30	0.03
Exciting	0.28	0.20	0.73	-0.21	-0.16
Similar to you	0.27	0.39	0.62	0.23	0.23
Passive	-0.10	-0.15	-0.79	0.12	0.25
¬ confident	-0.13	-0.05	-0.05	0.84	0.23
¬ Expert	-0.53	-0.23	-0.22	0.59	-0.11
¬ Friendly	-0.21	0.05	-0.21	0.14	0.82
¬ Competent	-0.53	-0.31	-0.01	0.16	0.61

Table 7: Five loaded factor labeled varimax

in the report, but I provided it as an attachment. Only NCL4 consisting of CL5, and CL6, and NCL3, consisting of CL11 and CL147 are tentative, due to R^2 close to .5. NCL2, and NCL1 also does not make sense, seince the R^2 is too small, close to zero, indicating that they are not homogeneous, at all. I tried to see something using 'proc tree' from this cluster, but nothing was visible even with 3 cluster, so probably to understand more we need to limit the observations to lower than 50, as you indicated in the class.

Second part To select segments we need to have small RMS STD, and SPRSQ, since these are measure of the similarity between groups. R^2 is also the indicator of homogeneity in the group. As a result we must have high R^2 and low SPRSQ. Cluster 9 shows to be elbow for the RMS STD. Although we can also find Cluster 5, as an elbow in R^2. Since the number of segments is prefered to be low, so I will select 5 clusters. Also on third cluster, remote access is an important matter.

Third part To find out the answer to this question I took the mean of all the id of each of the clusters. The result is shown in the following table. As could be seen the first cluster differentiator is high number of message usage, low monthly usage, and medium pda device price. The second cluster could be classified as group that needs multimedia and lightness. The third cluster has differentiator of low price of device, low monthly fee, who like to use messaging, and ergonomic is not important for them. The fourth cluster differentiator is innovation, but cell phone usage, messaging, and PIM is not that much important for them, although multimedia and ergonomic is important. Finally fifth cluster are willing to pay high price for the device, high monthly fee, but are probably business man who have high cell phone, PIM, and passive information usage. I would target fifth cluster, mean business man.

Fourth part The result of discriminant analysis procedure is shown in the following table. To

	1st Clus	2nd Clus	3rd Clus	4th Clus	Fifth Clus
Innovator	3	2	3	5	4
Use message	5	4	5	3	4
Use cell	6	5	6	4	6
Use PIM	4	4	5	3	5
Inf passive	5	4	4	5	7
INF active	5	5	4	5	4
remote access	4	4	5	3	7
Share info	4	4	3	5	7
Monitor	6	4	6	5	5
Email	5	5	5	6	3
Web	5	4	4	5	1
M media	4	4	4	4	1
ergonomic	5	5	4	5	4
monthly	25	33	27	35	40
price	304	436	134	597	790

Table 10: Five Cluster position in terms of segmentation variables

Variable	Communality
Likeable	0.72
Knowledgeable	0.75
Unattractive	0.81
Intelligent	0.56
Not similar to you	0.72
good looking	0.80
Unexciting	0.72
confident	0.78
friendly	0.79
believable	0.76
Expert	0.75
Ugly	0.78
Dont identify with	0.72
competent	0.78
active	0.73
irritating	0.67
Not trustworthy	0.86
dull	0.79
sincere	0.80

Table 8: Factor loading communality

Variable	Communality
Likeable	0.71765834
Knowledgeable	0.74771619
Unattractive	0.81194797
Intelligent	0.55651513
Not similar to you	0.72411968
good looking	0.79730854
Unexciting	0.71773715
confident	0.78233836
friendly	0.79245213
not believable	0.756696
Expert	0.7472156
Ugly	0.78024162
Dont identify with	0.71673012
competent	0.78207206
active	0.73251927
irritating	0.67460232
Not trustworthy	0.86443401
dull	0.78619071
not sincere	0.79502997

Table 9: Factor loading communality result of varimax

Variable	Can1	Can2	Can3	Can4
Age	0.73	0.18	0.39	0.54
Education	-0.94	0.10	-0.31	0.06
Income	-0.61	0.68	-0.04	-0.40
Construction	-0.26	-0.68	0.69	0.03
Emerg	-0.79	-0.30	-0.52	-0.09
Sales	0.89	0.29	-0.26	0.24
Service	0.87	-0.13	-0.47	-0.08
Professional	-0.85	0.19	0.10	-0.47
Compu	-0.87	0.46	0.17	0.05
PDA	-0.84	0.48	0.02	0.27
Cell_Ph	0.83	-0.38	-0.37	0.15
PC	0.98	0.00	0.19	-0.11
Away	0.98	-0.16	0.00	0.11
Bus_W	-0.82	0.57	-0.02	-0.11
PC_Mag	-0.81	0.15	-0.49	0.29
Field_S	-0.02	0.69	0.71	-0.14
M_Gourm	-0.01	0.76	0.18	0.63

Table 11: Between canonical structure

understand which of the demographic values are important first we need to look at the squared canonical correlation, which is like R^2 talking about how much of the variation is described by this canonical component. As the table shows the first canonical component is accounting for the highest variance, and second one's ration is half of the first. This suggests that we need to look at the highest values that are loaded into this canonical component. Age, sales, service, cell phone, PC, and remote workers are the one that are highly loaded into this component. While education, professional, computer, business week, and PC magazine are highly negatively loaded into it. On the second canonical component we have high load of field and stream, income, and gourment.

Fifth part To classify new customer I will plug in the attributes of the new customer into the canonical discriminant function and calculate score for each of the canonical discriminant component, and the one that has highest score will win.

Sixth part R-squre for the clustering is .93 showing that the clustering was able to separate

Variable	Can1	Can2	Can3	Can4
Age	0.14	0.05	0.14	0.38
Education	-0.38	0.07	-0.24	0.09
Income	-0.18	0.32	-0.02	-0.45
Construction	-0.13	-0.55	0.67	0.06
Emerg	-0.21	-0.13	-0.27	-0.10
Sales	0.30	0.16	-0.17	0.30
Service	0.28	-0.07	-0.29	-0.10
Professional	-0.16	0.06	0.04	-0.35
Compu	-0.27	0.22	0.10	0.05
PDA	-0.41	0.37	0.02	0.50
Cell_Ph	0.26	-0.19	-0.23	0.18
PC	0.31	0.00	0.12	-0.14
Away	0.37	-0.10	0.00	0.16
Bus_W	-0.13	0.15	-0.01	-0.07
PC_Mag	-0.24	0.07	-0.27	0.32
Field_S	-0.01	0.27	0.34	-0.13
M_Gourm	0.00	0.10	0.03	0.20

Table 12: Pooled within canonical structure

	Correlation	Correlation	Error	Correlation
1	0.535	0.447	0.057	0.286
2	0.369	0.221	0.069	0.136
3	0.310	0.211	0.072	0.096
4	0.164	-0.055	0.077	0.027

Table 13: squared canonical correlation

the variables, yet there was a warining saying that there is correlation between variables. Cluster means are shown in the following table. In comparison to hierarchical cluster we have difference R-squre for four clusters. For four clusters we have .5 R^2 for non hierarchical clustering, but here we have much higher R^2. Regarding the mean of the cluster we also have different values, but not dramatically different from hierarchical clustering method.

Cluster	Innovator	Use message	Use cell	Use PIM	Inf passive
1	4	4	5	3	5
2	4	4	6	5	7
3	3	5	6	4	4
4	4	4	6	4	4
Cluster	INF active	remote access	Share info	Monitor	Email
1	5	4	4	4	4
2	4	7	7	5	3
3	4	4	3	5	5
4	4	4	4	5	5
Cluster	Web	M media	ergonomic	monthly	price
1	4	4	5	36.50	498.50
2	1	1	4	40.00	790.00
3	4	4	4	24.31	208.24
4	5	4	5	29.43	361.14

Table 14: Non hierarchical clustering result