不同脂肪肝程度對國人健康狀態影響

第一組
統計所　張麒仙
動科所　張馨文
生工所　凌家宜
生工所　張少華
統計所　李冠薇

目錄

- 背景介紹
- 資料來源
- 資料介紹
- 分析方法與結果
- 參考文獻

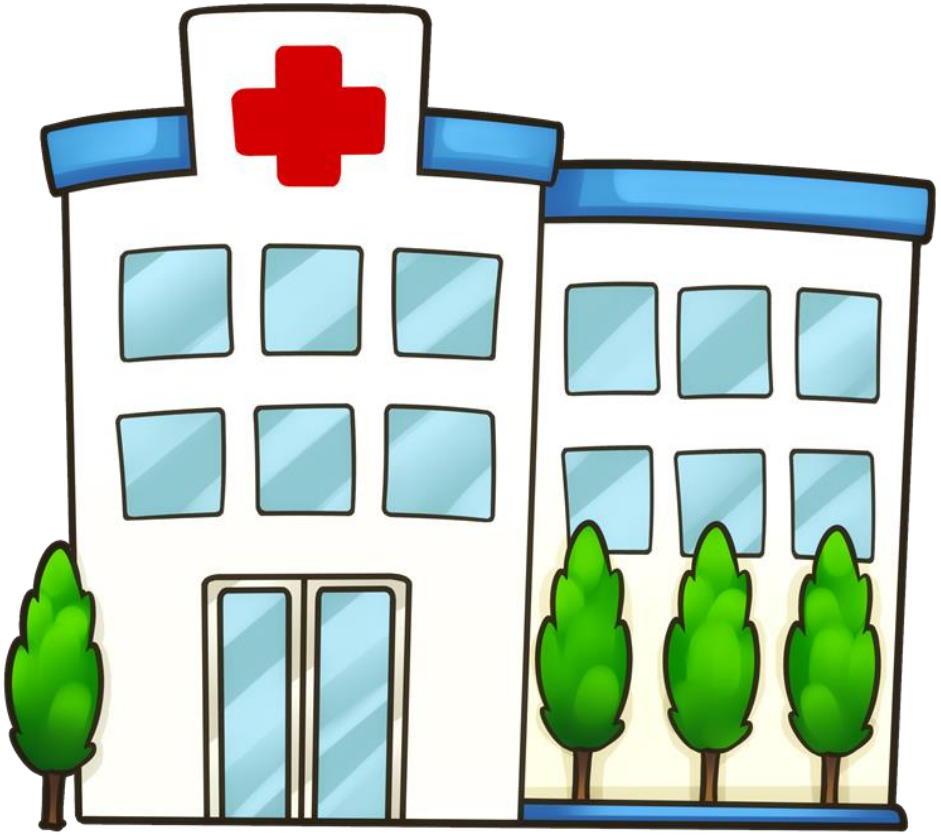
現代人因不適當的飲食習慣與運動量不足等因素
導致三高（高血脂、 高血糖、高血壓）等現代文明病
就連負責身體代謝最重要的器官之一「肝臟」
也躲不過變胖的厄運，變成了「脂肪肝」

- ✓ 沒有酗酒習慣的病人，亦有可能會有脂肪肝的發生，稱為非酒精性脂肪肝疾病
- ✓ 脂肪肝過去被認為是良性可逆的疾病，因此受重視程度較低
- ✓ 近年來的研究發現，脂肪肝可能會進一步的演變成肝炎、肝硬化、肝細胞癌
- ✓ 故我們有需要更加了解它


正常肝臟
脂肪肝
肝硬化

資料來源

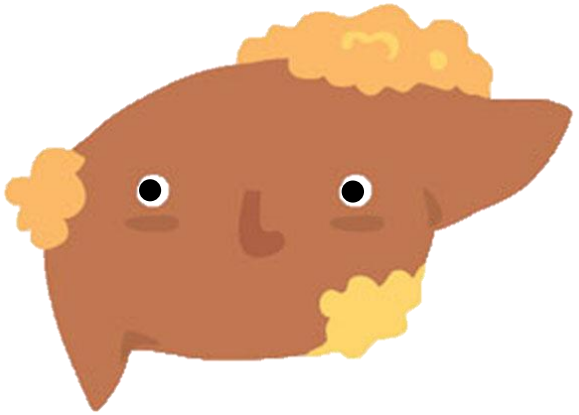
2012年馬偕醫院高價健檢結果資料



馬偕紀念醫院
MacKay Memorial Hospital

資料介紹

- 脂肪肝就是肝細胞內有脂肪聚積
- 也就是肝臟比較『油』



肝內脂肪積聚	
超過肝重的5%	輕度脂肪肝
超過肝重的10%	中度脂肪肝
超過肝重的25%	重度脂肪肝

資料介紹

y (脂肪肝)	情況
0	正常
1	輕度脂肪肝
2	中度以上脂肪肝

- 資料量：受測者323位

變數選擇

- 區別變數：

從資料完整的46個變數選出23個Anova分析結果顯示組間差異顯著的

再做Manova分析確認仍然顯著

變數	Pr > F	變數	Pr > F	變數	Pr > F
SGOT_AST	0.0756	waist	<.0001	Segment	0.1948
SGPT_ALT	0.0835	Buttock	<.0001	Eosinophil	0.3168
Height	0.3614	GlucoseAC	<.0001	Basophil	0.142
Weight	<.0001	T_Cholesterol	0.0043	Monocyte	0.1273
BMI	<.0001	Triglyceride	0.001	Lymphocyte	0.2349
IBW_L	0.3548	UricAcid	<.0001	Platelet	0.1071
IBW_U	0.3546	Creatinine	0.0693	Specific_Gravity	0.7347
BP_H	<.0001	HDL	<.0001	pH	0.0884
BP_L	<.0001	Hb	<.0001	Urobilinogen	0.1097
PulseRate	0.3639	RBC	0.0345	RBC_1	0.5052
Temp	<.0001	Ht	<.0001	WBC_1	0.4853
Sex	0.0039	MCV	0.0269	EpithelialCell	0.0495
Age	<.0001	MCH	0.0113	nAFP	0.4495
BodyFat	<.0001	MCHC_1	0.0048	nBodyFat	<.0001
IBF_L	0.4796	MCHC_2	0.0048	nCEA	0.0732
IBF_U	0.2108	WBCcount	0.0074		

Multivariate Statistics and Exact F Statistics

S=1 M=9.5 N=149.5

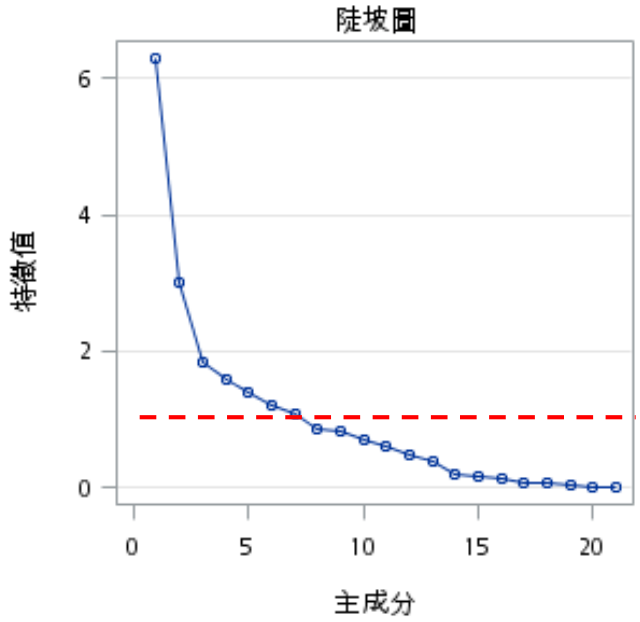
統計值	值	F 值	分子自由度	分母自由度	Pr > F
Wilks' Lambda	0.61433947	9.00	21	301	<.0001
Pillai's Trace	0.38566053	9.00	21	301	<.0001
Hotelling-Lawley Trace	0.62776453	9.00	21	301	<.0001
Roy's Greatest Root	0.62776453	9.00	21	301	<.0001

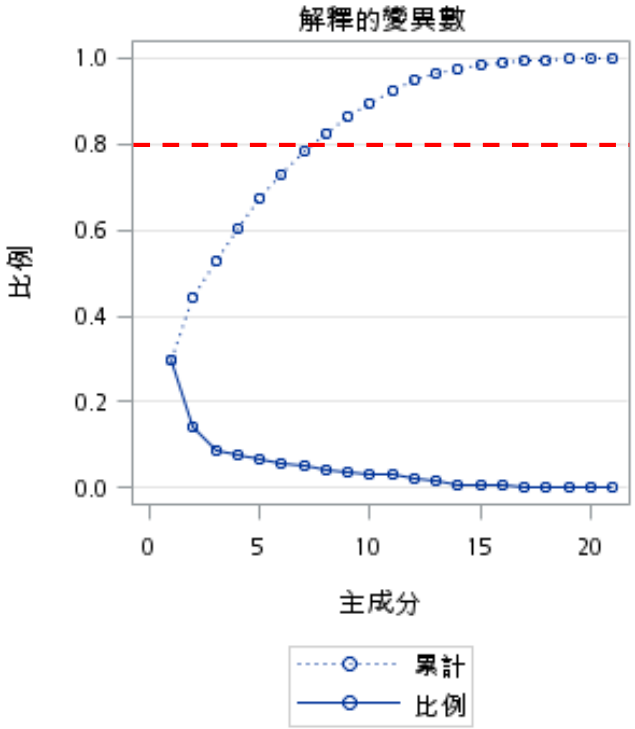
分析方法

- ✓因素分析
- ✓集群分析
- ✓判別分析
- ✓羅吉斯分析
- ✓典型相關分析

主成分分析

Principal Component Analysis

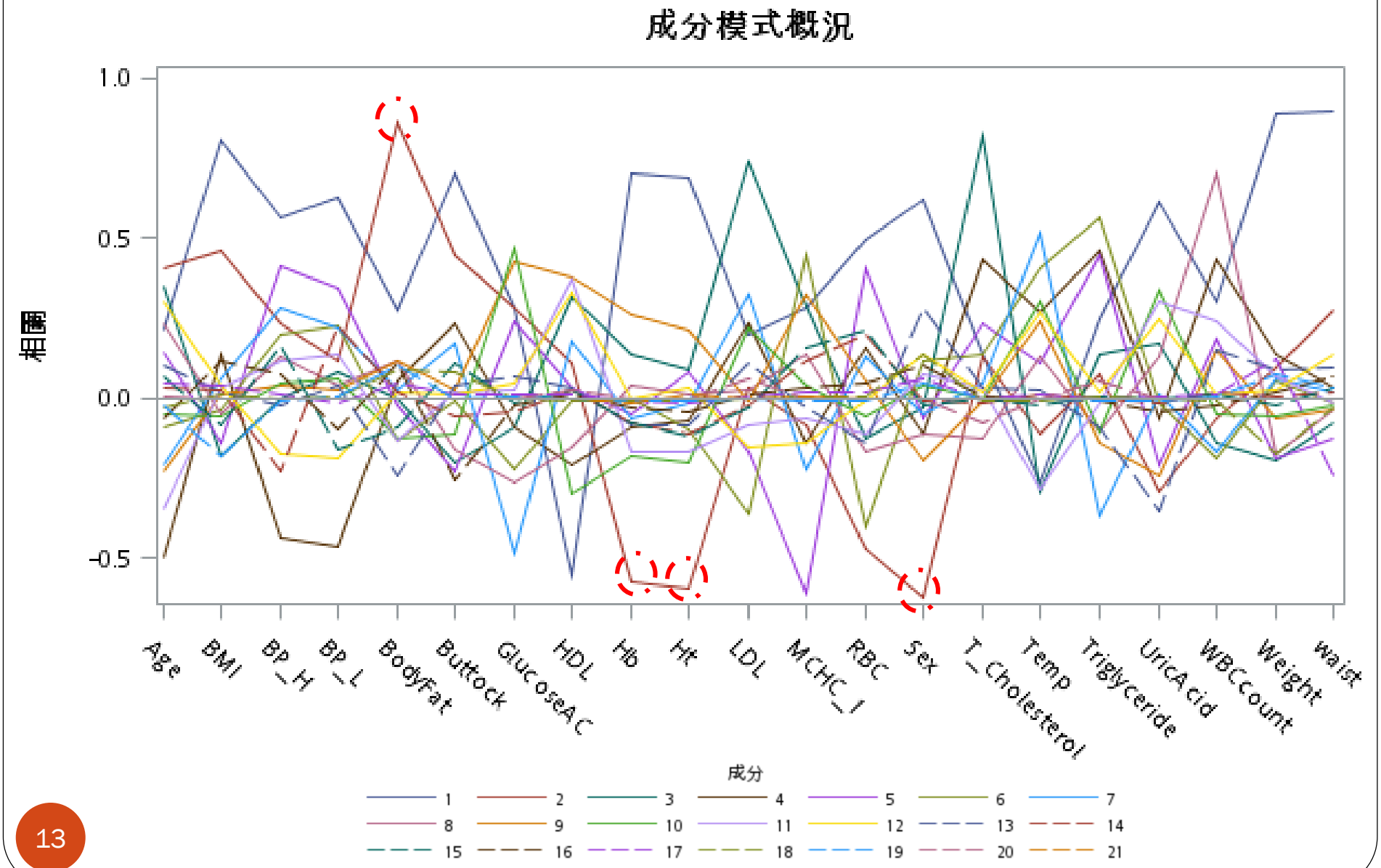




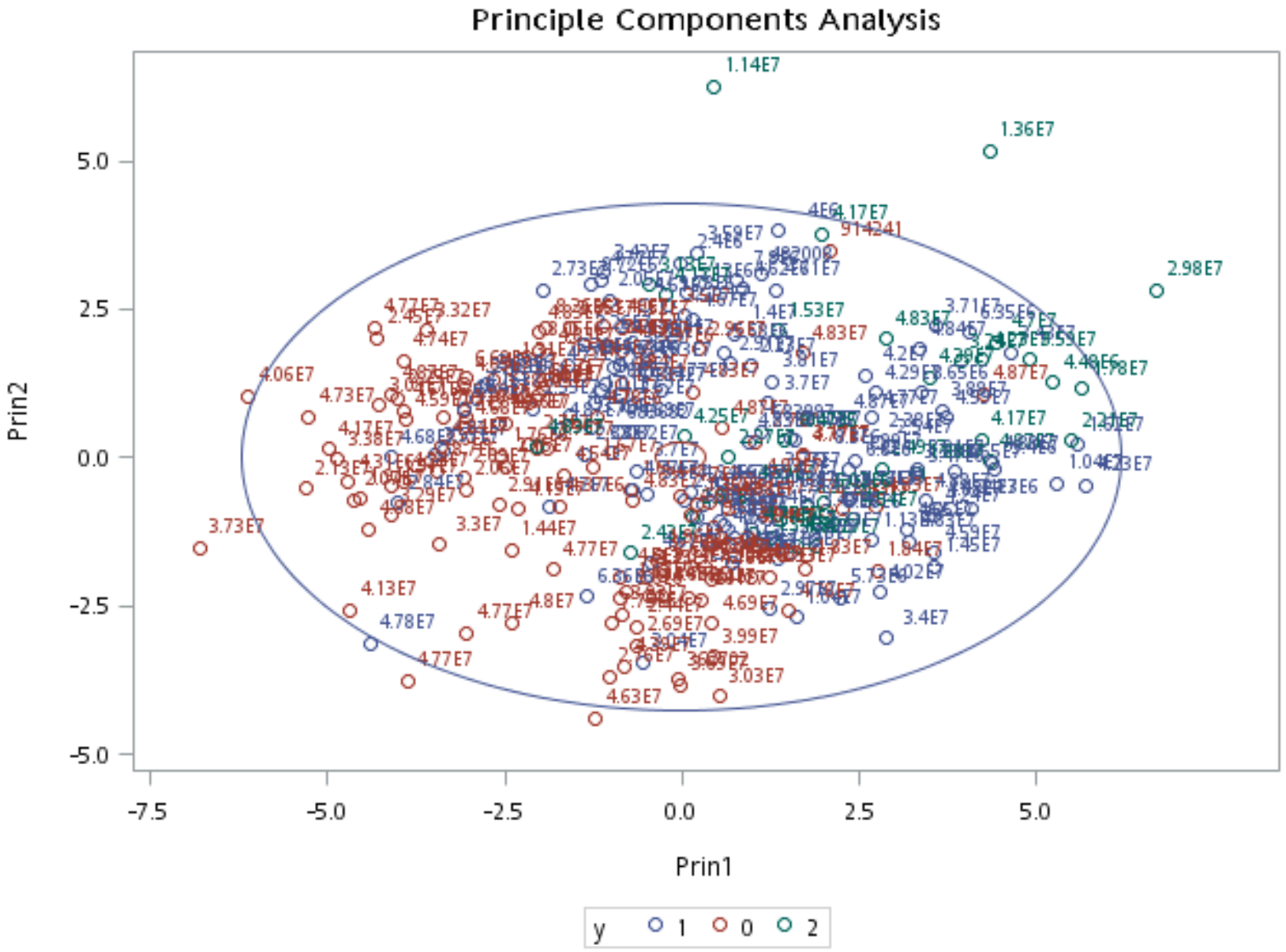
相關矩陣的特徵值

	特徵值	差異	比例	累計
1	6.29143181	3.29835479	0.2996	0.2996
2	2.99307702	1.14111511	0.1425	0.4421
3	1.85196192	0.26967651	0.0882	0.5303
4	1.58228540	0.17400646	0.0753	0.6057
5	1.40827895	0.19129378	0.0671	0.6727
6	1.21698516	0.13119503	0.0580	0.7307
7	1.08579013	0.23316408	0.0517	0.7824
8	0.85262605	0.03367653	0.0406	0.8230

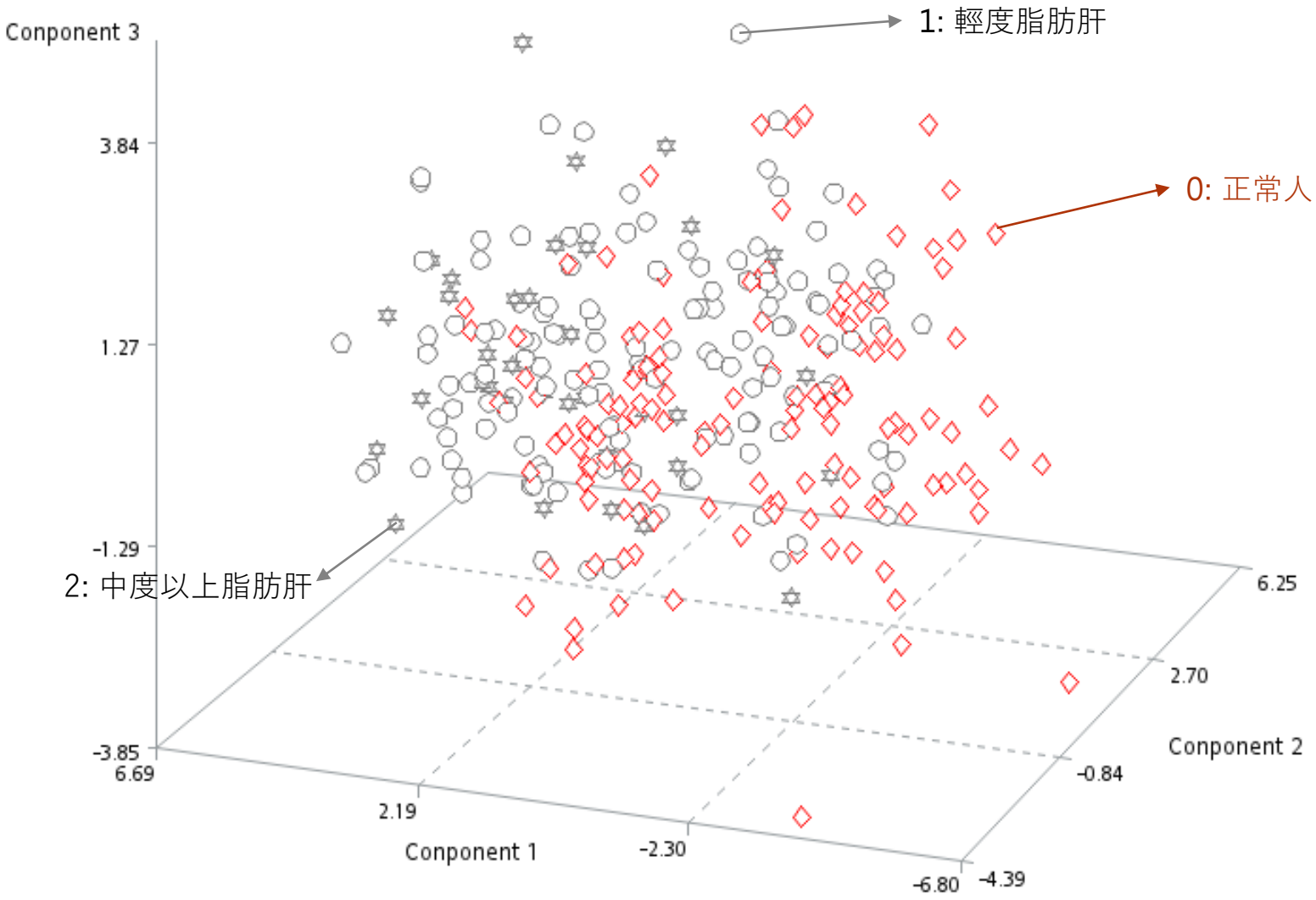
Principal Component Analysis



Principal Component Analysis



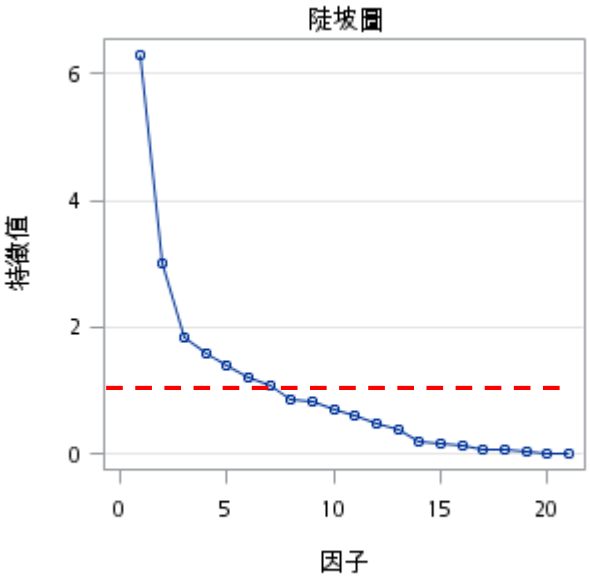
Principal Component Analysis

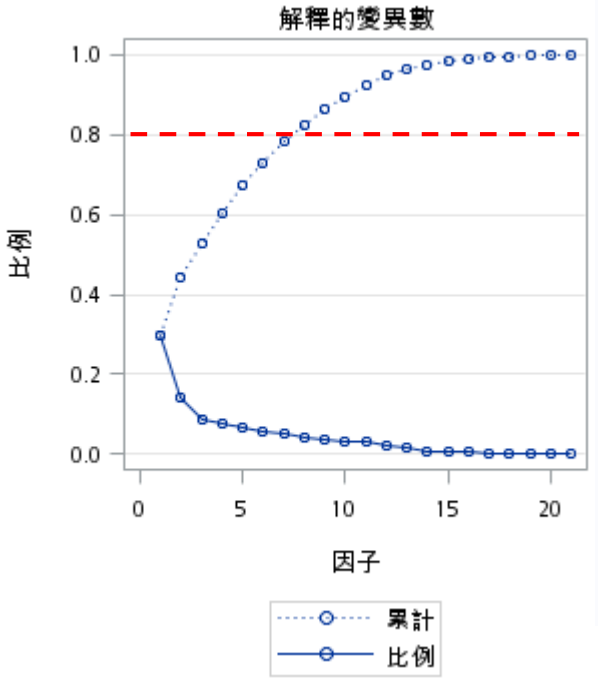


因素分析

Exploratory Factor Analysis:

Principal Components method





相關矩陣的特徵值: 總計 = 21 平均 = 1

	特徵值	差異	比例	累計
1	6.29143181	3.29835479	0.2996	0.2996
2	2.99307702	1.14111511	0.1425	0.4421
3	1.85196192	0.26967651	0.0882	0.5303
4	1.58228540	0.17400646	0.0753	0.6057
5	1.40827895	0.19129378	0.0671	0.6727
6	1.21698516	0.13119503	0.0580	0.7307
7	1.08579013	0.23316408	0.0517	0.7824
8	0.85262605	0.03367653	0.0406	0.8230

Exploratory Factor Analysis:

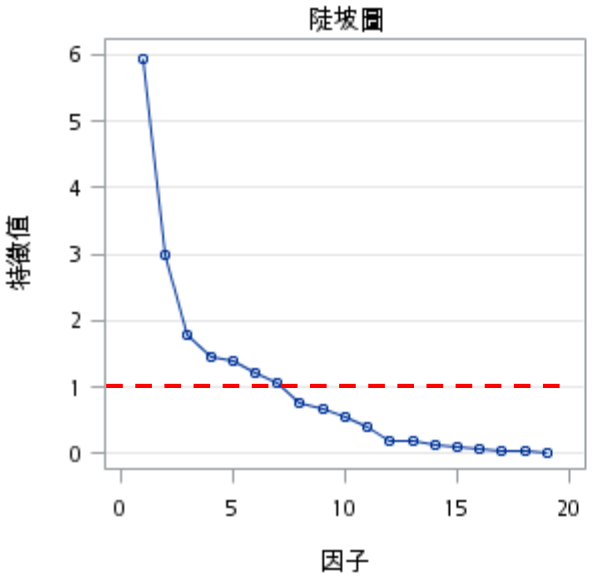
Principal Components method

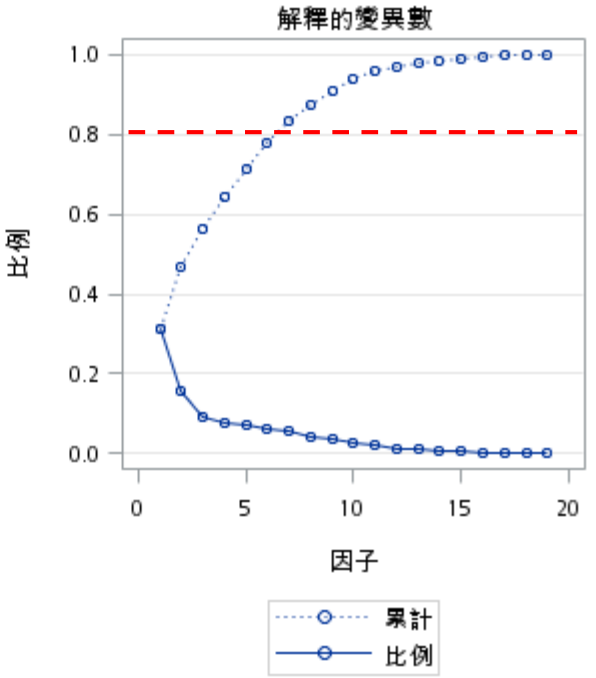
最終公因子變異數估計值: 總計 = 16.429810

Age	BMI	BP_H	BP_L	BodyFat
0.654287	0.9372463	0.852709	0.84894297	0.841725
Buttock	GlucoseAC	HDL	Hb	Ht
0.869951	0.51648091	0.496422	0.84991093	0.852039
LDL	MCHC_1	RBC	Sex	waist
0.905103	0.80752973	0.849662	0.80838068	0.903295
Temp	Triglyceride	UricAcid	WBCcount	Weight
0.6795	0.94914908	0.534361	0.39986035	0.897502
T_Cholesterol				
0.97575396				

Exploratory Factor Analysis:

Principal Components method





相關矩陣的特徵值: 總計 = 19 平均 = 1				
	特徵值	差異	比例	累計
1	5.94960535	2.96882369	0.3131	0.3131
2	2.98078166	1.19138171	0.1569	0.4700
3	1.78939995	0.32098752	0.0942	0.5642
4	1.46841243	0.08584137	0.0773	0.6415
5	1.38257106	0.18223337	0.0728	0.7143
6	1.20033769	0.14111639	0.0632	0.7774
7	1.05922129	0.28620793	0.0557	0.8332

Exploratory Factor Analysis:

Principal Components method

透過每個因子所解釋的變異數						
Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
5.9496053	2.9807817	1.7893999	1.4684124	1.3825711	1.2003377	1.0592213

Kaiser 取樣適當性量數: 整體 MSA = 0.68070232

最終公因子變異數估計值: 總計 = 15.830329

Age	BMI	BP_H	BP_L
0.70184137	0.94340941	0.869781	0.84403118
Buttock	**GlucoseAC**	**Hb**	**Ht**
0.8951698	0.58395157	0.858738	0.85805339
MCHC_1	**RBC**	**Sex**	**T_Cholesterol**
0.84106926	0.88292686	0.804611	0.97321043
Triglyceride	**UricAcid**	**Weight**	**waist**
0.94638125	0.54095201	0.918746	0.9061845
BodyFat	**LDL**	**Temp**	
0.83776215	0.93450454	0.689006	

因子模型								
		Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
Age	Age	0.24077	0.38679	0.26514	-0.63839	-0.10793	0.00328	-0.06909
BMI	BMI	0.80348	0.45417	-0.16272	0.24202	-0.05832	-0.05410	0.01347
BP_H	BP_H	0.58597	0.21323	-0.12102	-0.49000	0.27087	0.19030	0.34150
BP_L	BP_L	0.65422	0.09148	-0.05272	-0.48667	0.21002	0.22529	0.27049
BodyFat	BodyFat	0.28719	0.85959	0.01704	0.02893	-0.03591	-0.11461	0.02881
Buttock	Buttock	0.70429	0.44210	-0.16933	0.39900	-0.07270	-0.06222	0.08162
GlucoseAC	GlucoseAC	0.28812	0.28156	-0.10266	-0.18501	0.15818	-0.19755	-0.55933
Hb	Hb	0.69918	-0.59149	0.07319	-0.05374	-0.07650	0.02658	-0.07227
Ht	Ht	0.68328	-0.61064	0.03021	-0.07244	0.04332	-0.08839	-0.04951
LDL	LDL	0.21678	0.01929	0.77315	0.11584	-0.14870	-0.38147	0.32912
MCHC_1	MCHC_1	0.29627	-0.09853	0.24602	0.06994	-0.59344	0.55687	-0.12604
RBC	RBC	0.47835	-0.47235	-0.11828	0.03649	0.41959	-0.48938	0.01124
Sex	Sex	0.61254	-0.63752	-0.07696	-0.01926	-0.04572	0.10790	-0.05421
T_Cholesterol	T_Cholesterol	0.16584	0.11159	0.87961	0.18280	0.34943	0.06214	0.01253
Temp	Temp	-0.27968	-0.09690	-0.23266	0.32648	0.28572	0.30709	0.51452
Triglyceride	Triglyceride	0.23114	0.08094	0.24348	0.23987	0.62478	0.48972	-0.37337
UricAcid	UricAcid	0.60807	-0.30289	0.14383	-0.01236	-0.23076	0.02413	0.06918
Weight	Weight	0.88460	0.08658	-0.18794	0.29915	-0.06067	0.01197	-0.00991
waist	waist	0.89846	0.25972	-0.08989	0.12550	-0.08522	-0.01994	-0.00318

Exploratory Factor Analysis:

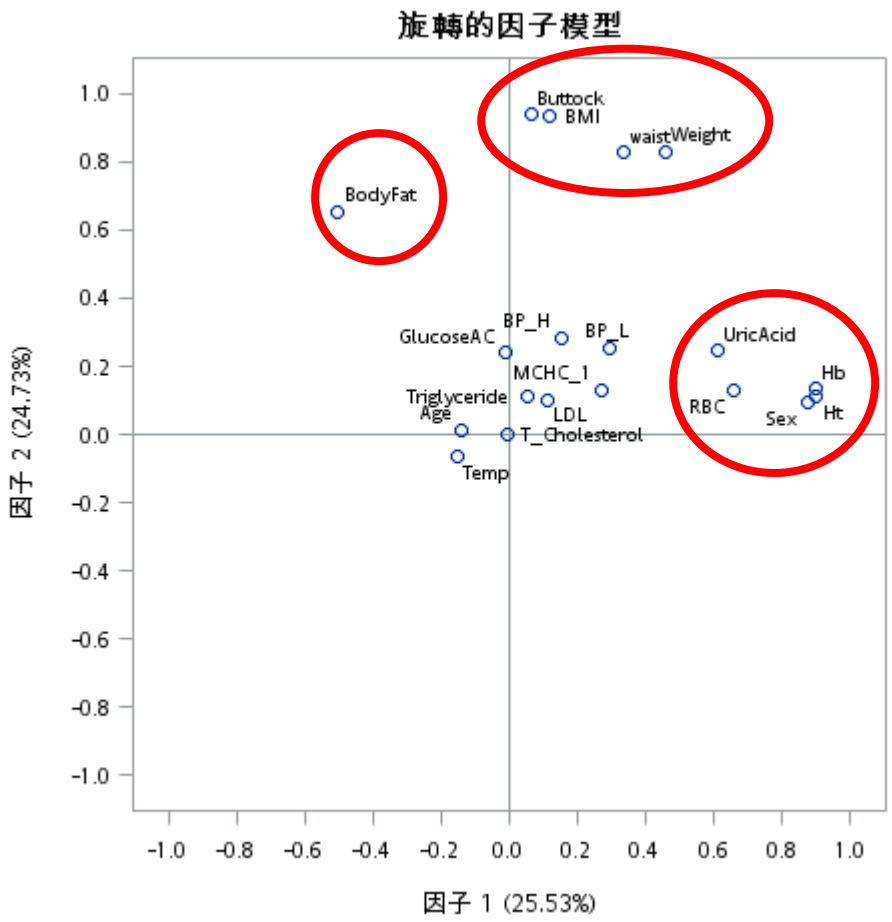
Principal Components method
(varimax rotation)

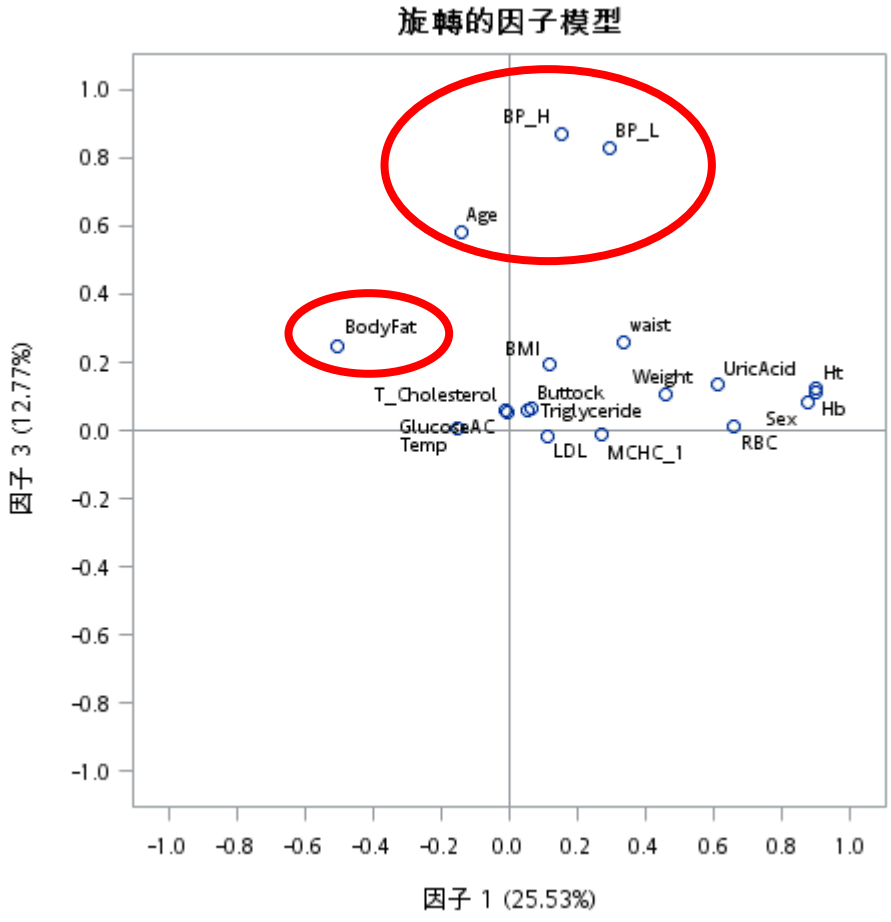
透過每個因子所解釋的變異數						
Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
4.0410247	3.9147085	2.0221923	1.7120845	1.4380382	1.3714237	1.3308576

旋轉的因子模型								
		Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
Age	Age	-0.14002	0.01111	0.58147	0.18540	-0.52323	-0.07778	0.17265
BMI	BMI	0.12076	0.93345	0.19215	0.02180	-0.13118	0.05129	0.01605
BP_H	BP_H	0.15216	0.28228	0.86863	-0.04314	0.01483	0.05374	-0.08632
BP_L	BP_L	0.29610	0.25148	0.82803	-0.01535	-0.02919	0.07995	-0.00020
BodyFat	BodyFat	-0.50695	0.65325	0.24976	0.14955	-0.26317	-0.00496	-0.00057
Buttock	Buttock	0.06387	0.94002	0.06224	0.04796	0.00449	0.03457	0.00817
GlucoseAC	GlucoseAC	-0.00943	0.24337	0.06076	-0.18289	-0.63787	0.21180	-0.18910
Hb	Hb	0.89859	0.13402	0.11068	0.06280	-0.07935	0.04037	0.09592
Ht	Ht	0.90185	0.11415	0.12165	0.05972	-0.07732	0.03484	-0.07826
LDL	LDL	0.11170	0.09824	-0.01571	0.94041	-0.03360	-0.16283	0.01067
MCHC_1	MCHC_1	0.27003	0.13015	-0.01397	0.05145	-0.03360	0.04252	0.86339
RBC	RBC	0.66003	0.12978	0.00991	0.06406	-0.03535	0.03340	-0.65106
Sex	Sex	0.87676	0.09365	0.08245	-0.09675	0.03334	0.04970	0.08600
T_Cholesterol	T_Cholesterol	-0.00749	0.00094	0.05138	0.80749	-0.03542	0.56223	0.03354
Temp	Temp	-0.15479	-0.06216	0.00425	-0.13396	0.78725	0.10944	-0.10712
Triglyceride	Triglyceride	0.05347	0.11267	0.05857	0.00583	-0.00194	0.96280	0.01922
UricAcid	UricAcid	0.61098	0.24596	0.13287	0.18655	-0.04392	-0.08675	0.21271
Weight	Weight	0.45970	0.82763	0.10345	-0.03628	-0.02620	0.08780	0.04518
waist	waist	0.33587	0.82796	0.26119	0.05386	-0.17115	0.04832	0.07152

Exploratory Factor Analysis:

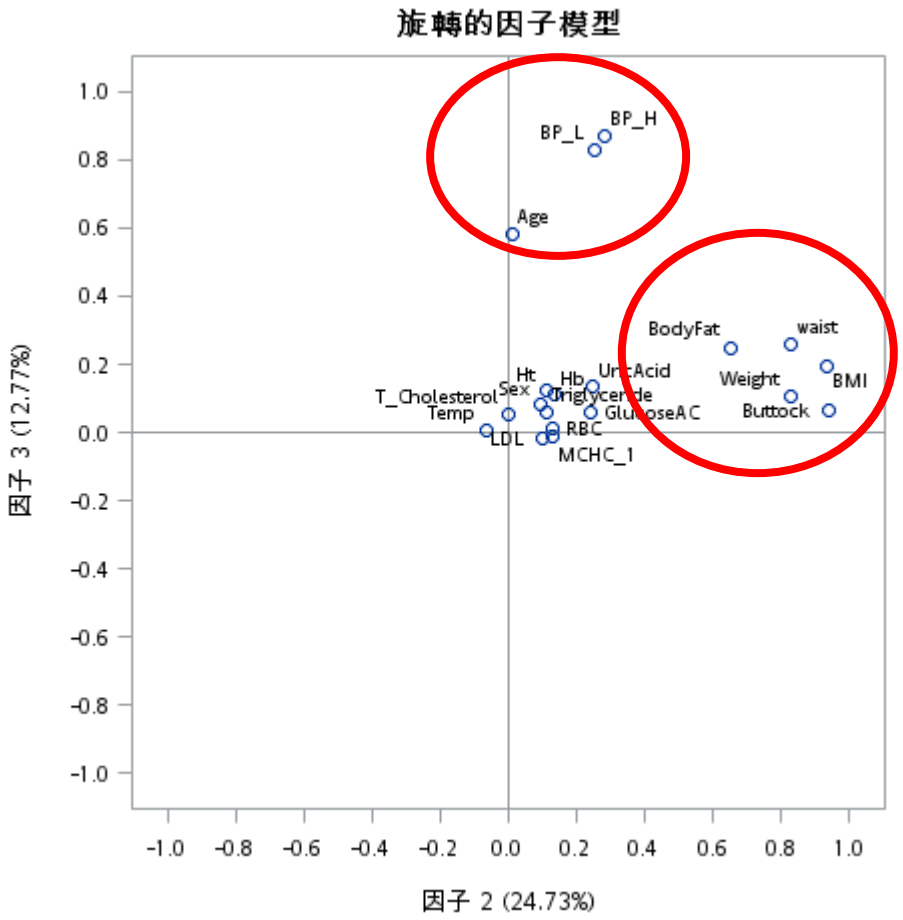
Principal Components method(varimax rotation)

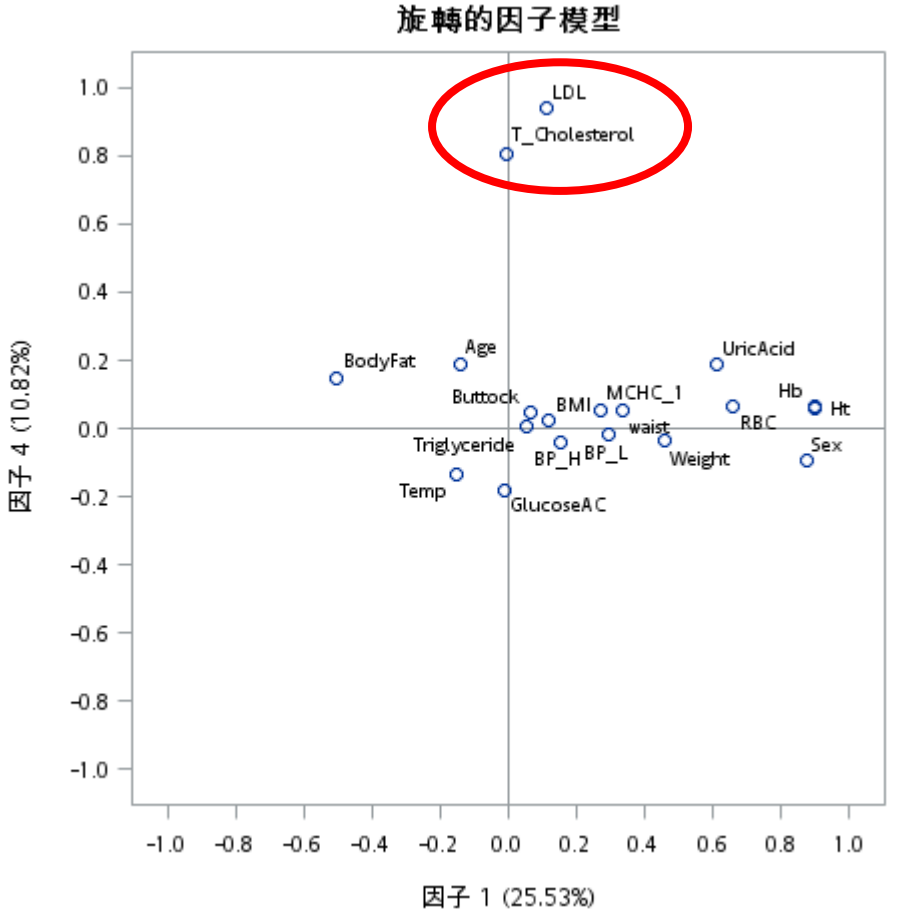




Exploratory Factor Analysis:

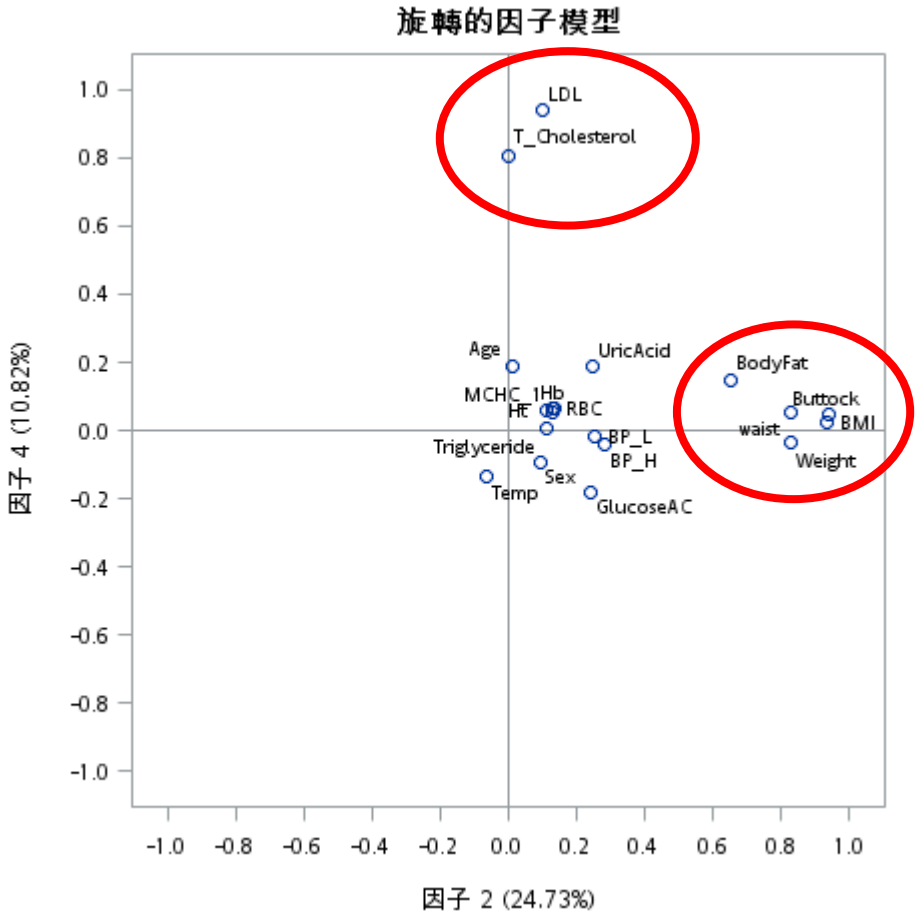
Principal Components method(varimax rotation)

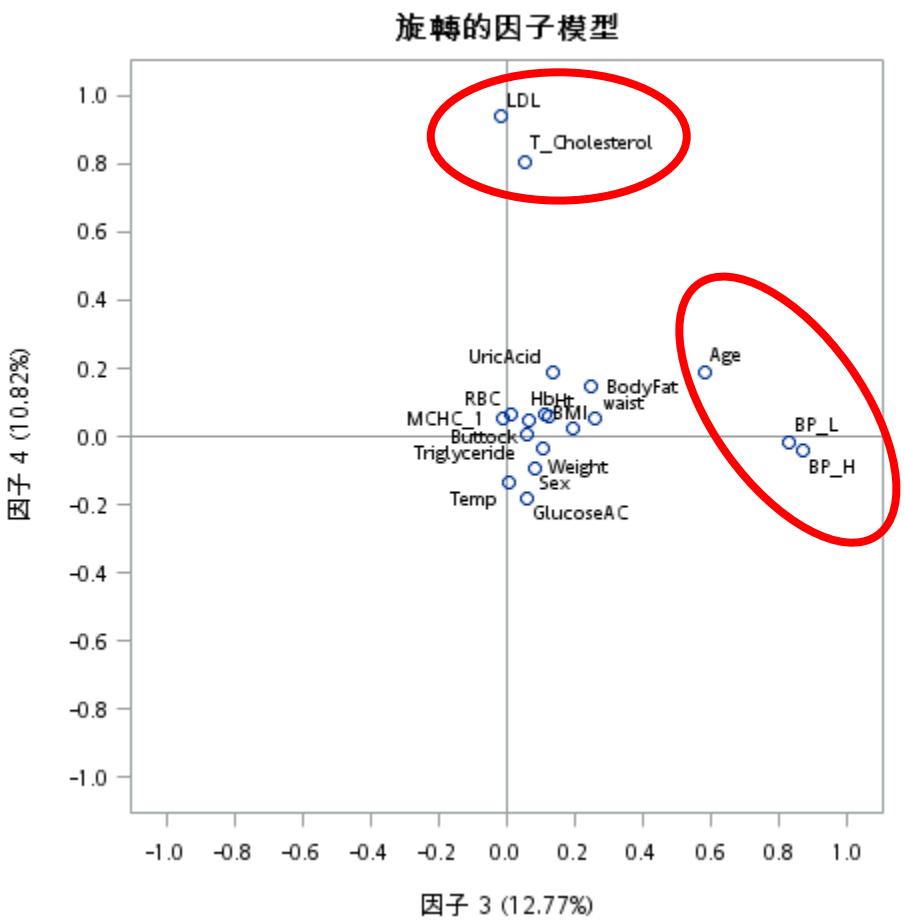




Exploratory Factor Analysis:

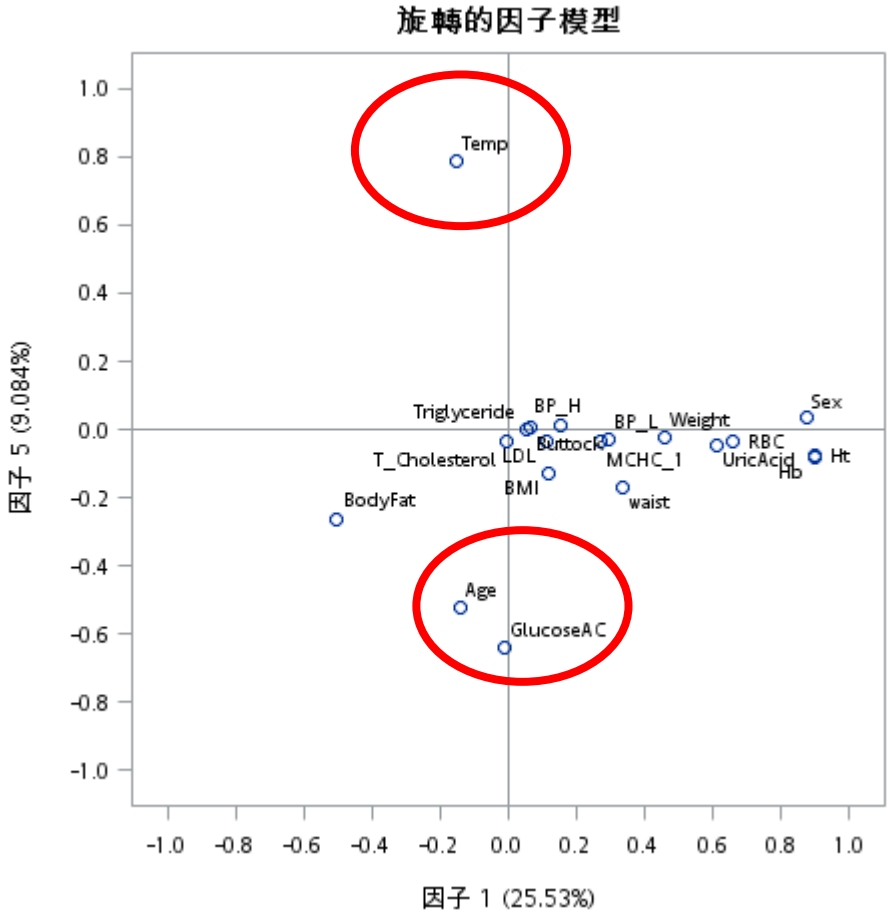
Principal Components method(varimax rotation)

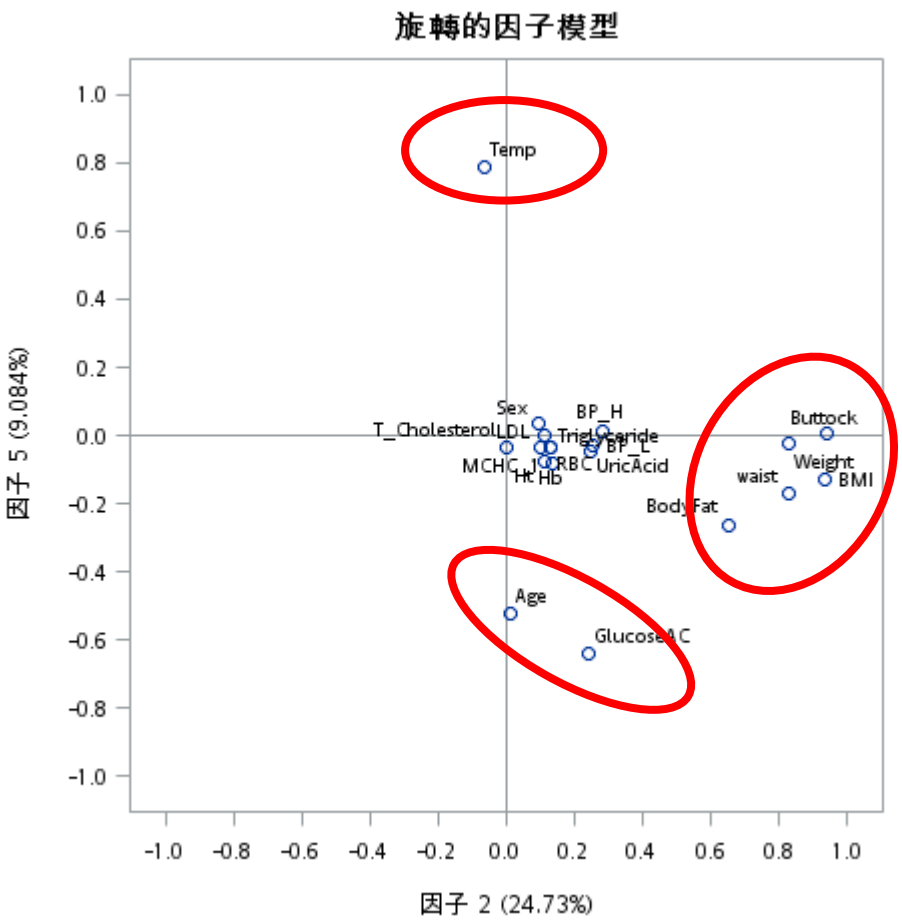




Exploratory Factor Analysis:

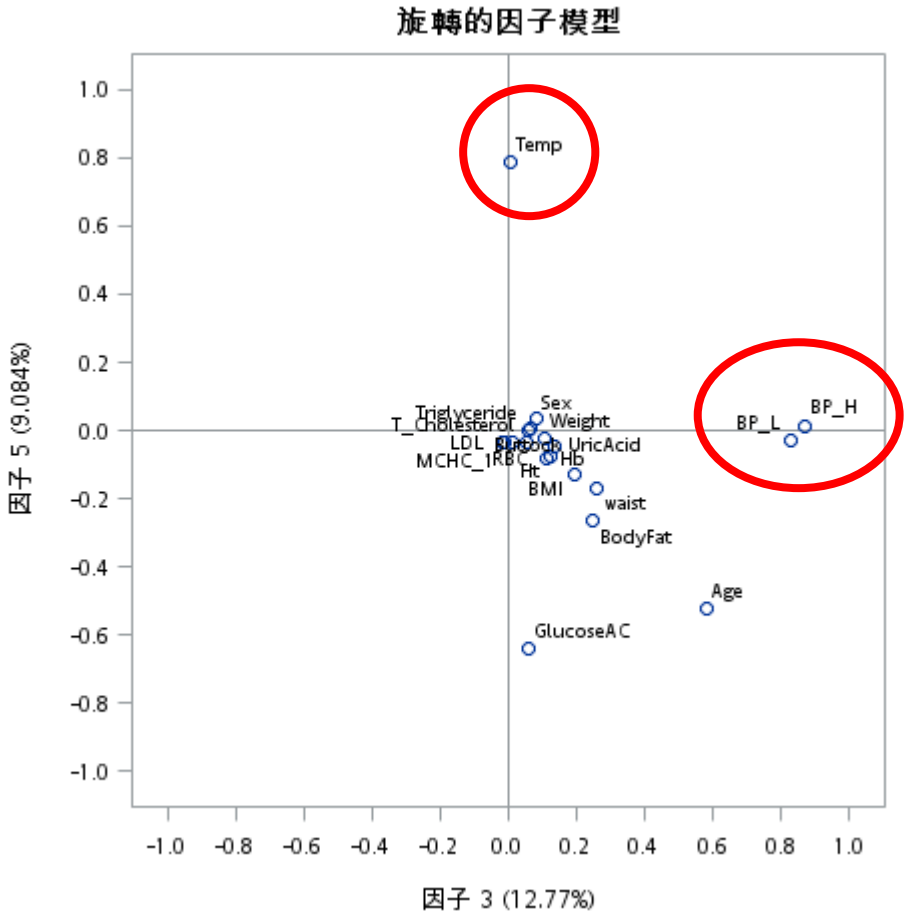
Principal Components method(varimax rotation)

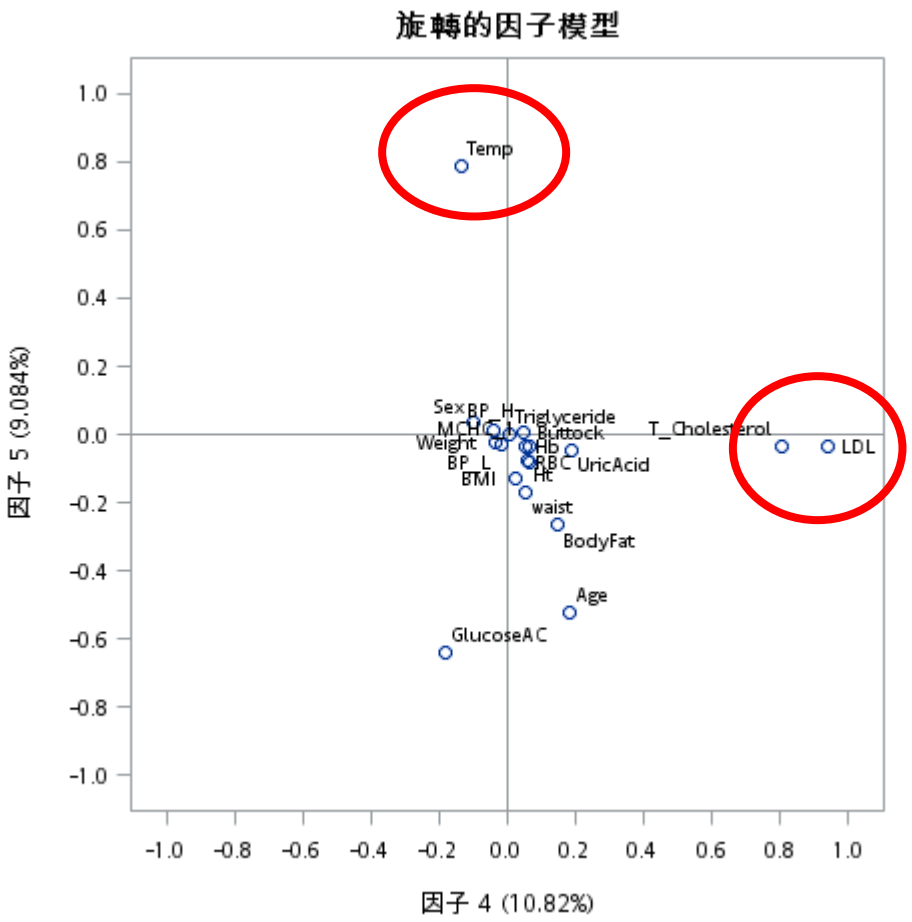




Exploratory Factor Analysis:

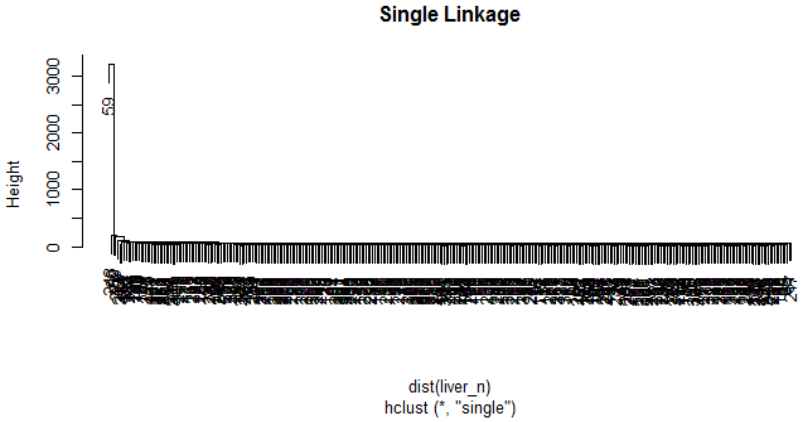
Principal Components method(varimax rotation)

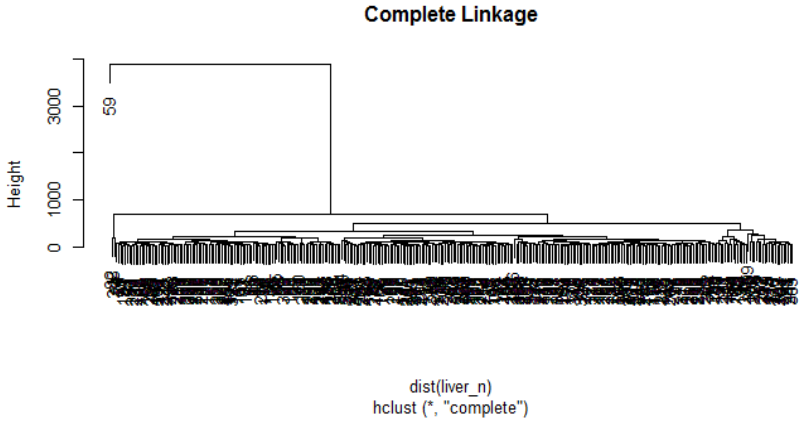


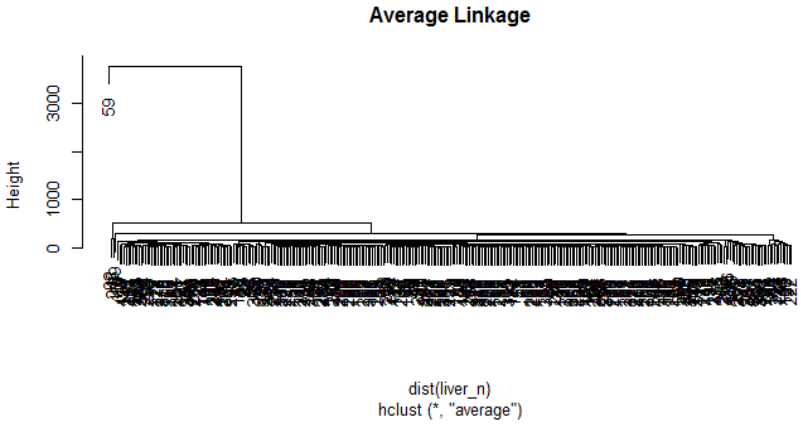


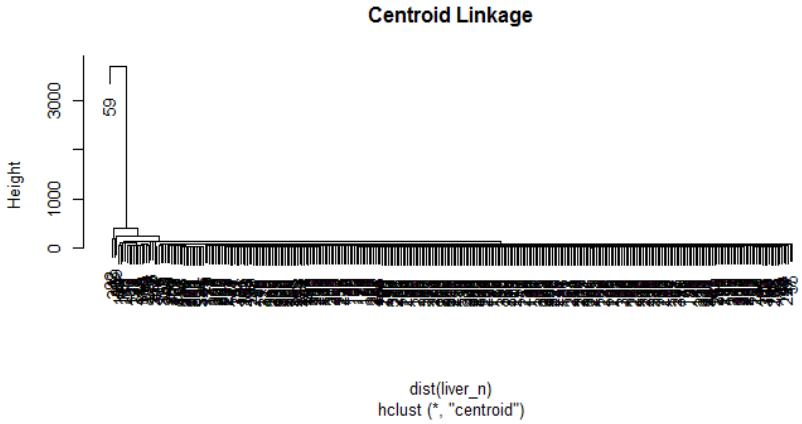
集群分析

將受試者進行分群

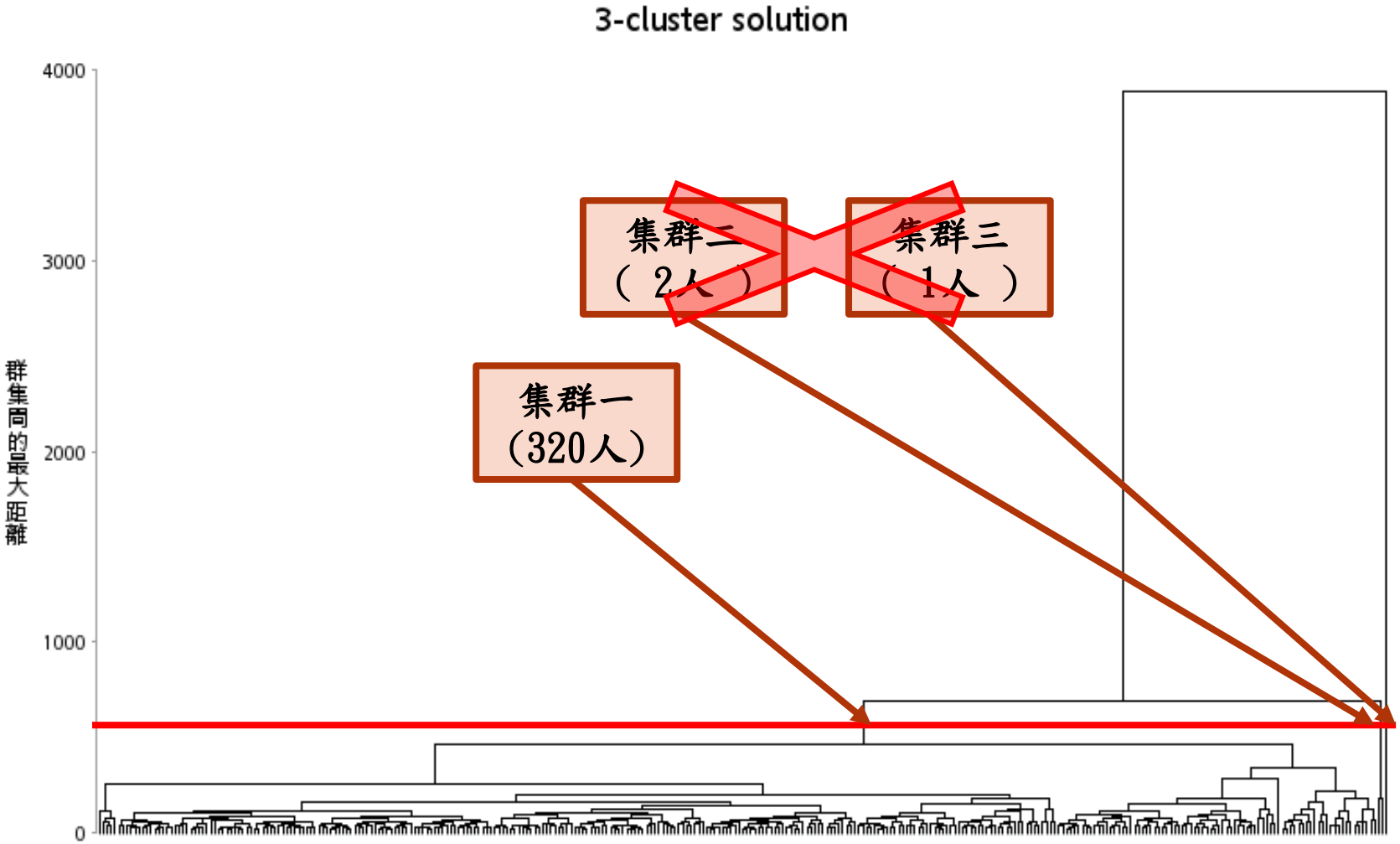

Single Linkage
Height
0
1000
2000
3000
59
dist(liver_n)
hclust (*, "single")


Complete Linkage
Height
0
1000
3000
59
dist(liver_n)
hclust (*, "complete")


Average Linkage
Height
0
1000
3000
59
dist(liver_n)
hclust (*, "average")

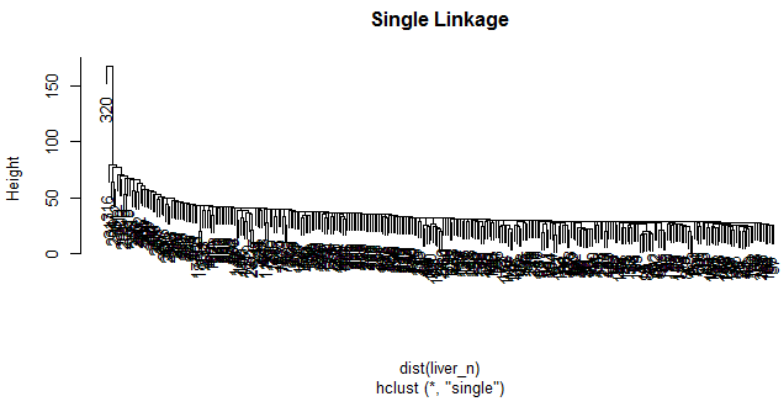

Centroid Linkage
Height
0
1000
3000
59
dist(liver_n)
hclust (*, "centroid")

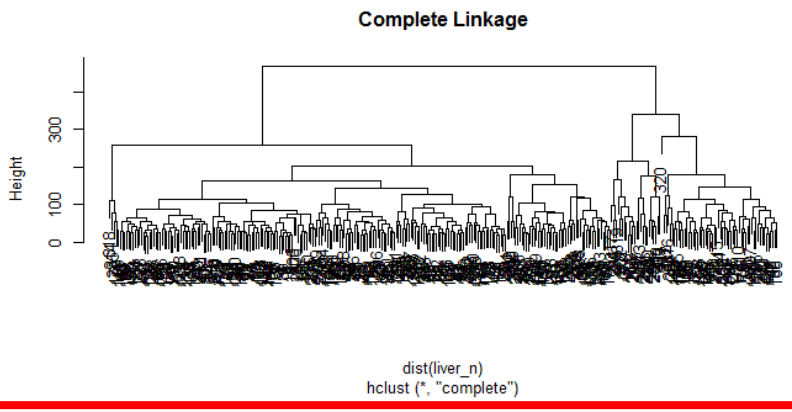
利用完整連結進行分群

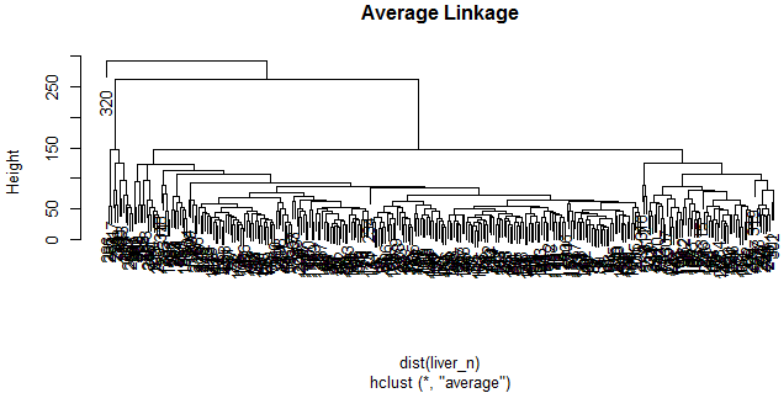

3-cluster solution
4000
3000
2000
1000
0
群集間的最大距離
集群二
（2人）
集群三
（1人）
集群一
(320人)

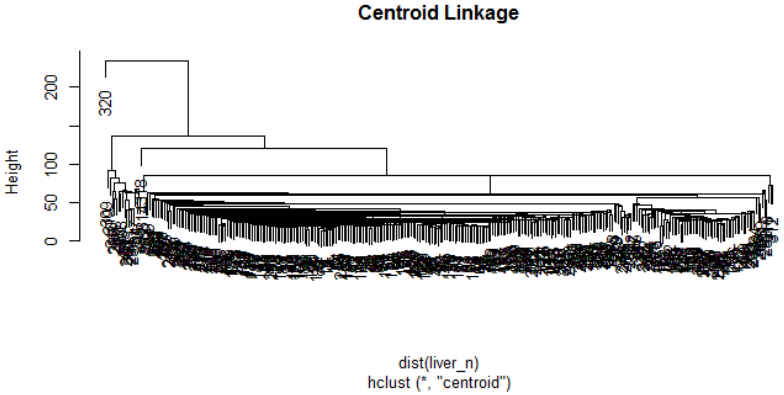
階層式集群分析

將受試者(刪除3人)進行分群

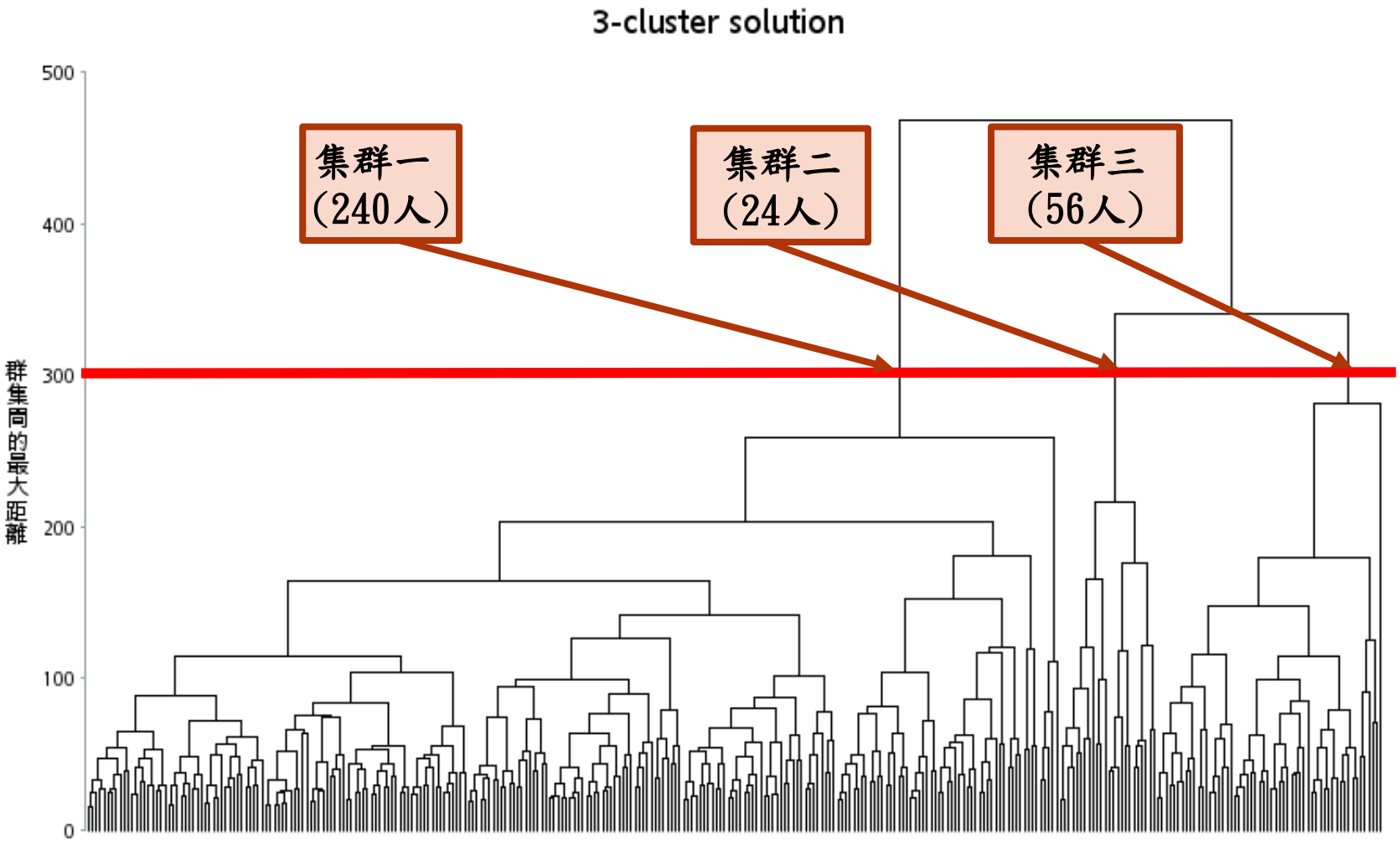






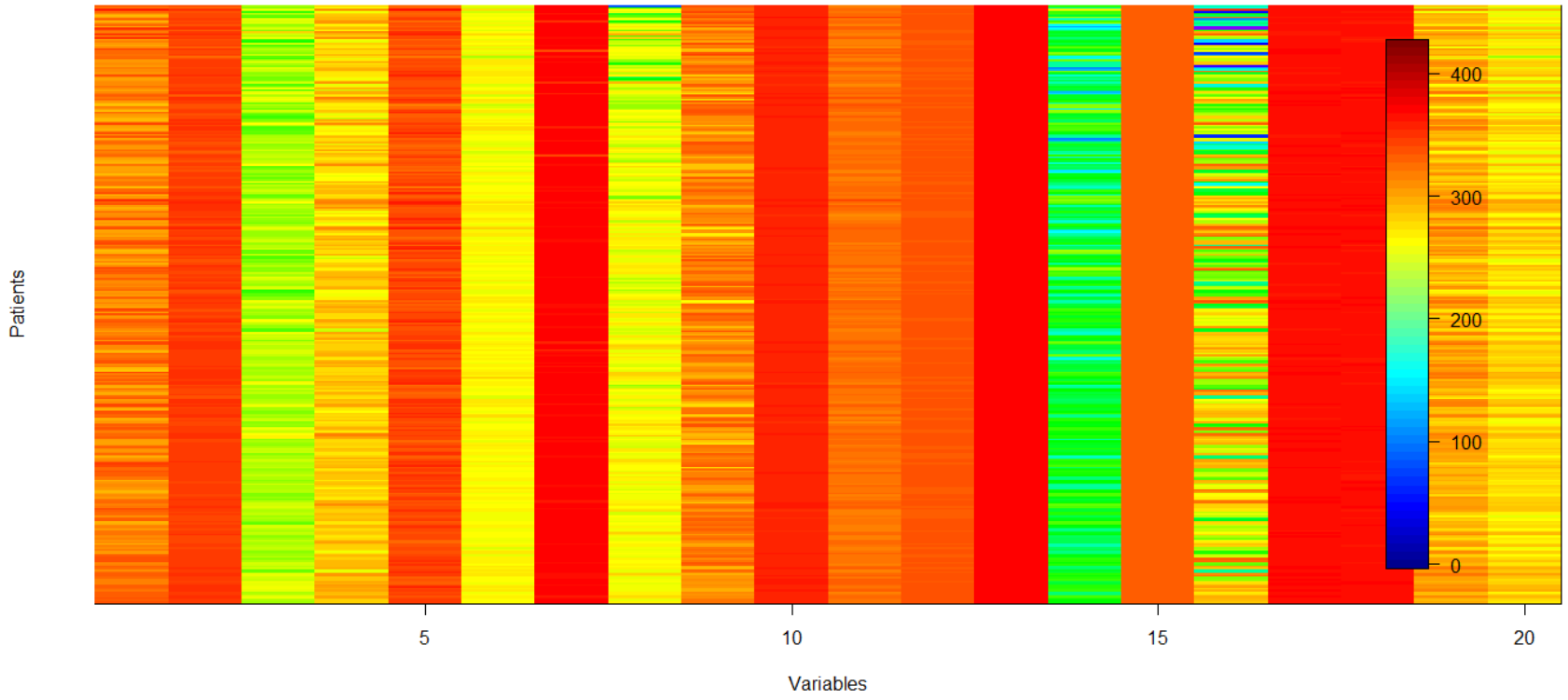


利用完整連結進行分群


3-cluster solution
集群一
(240人)
集群二
(24人)
集群三
(56人)
群集間的最大距離
500
400
300
200
100
0

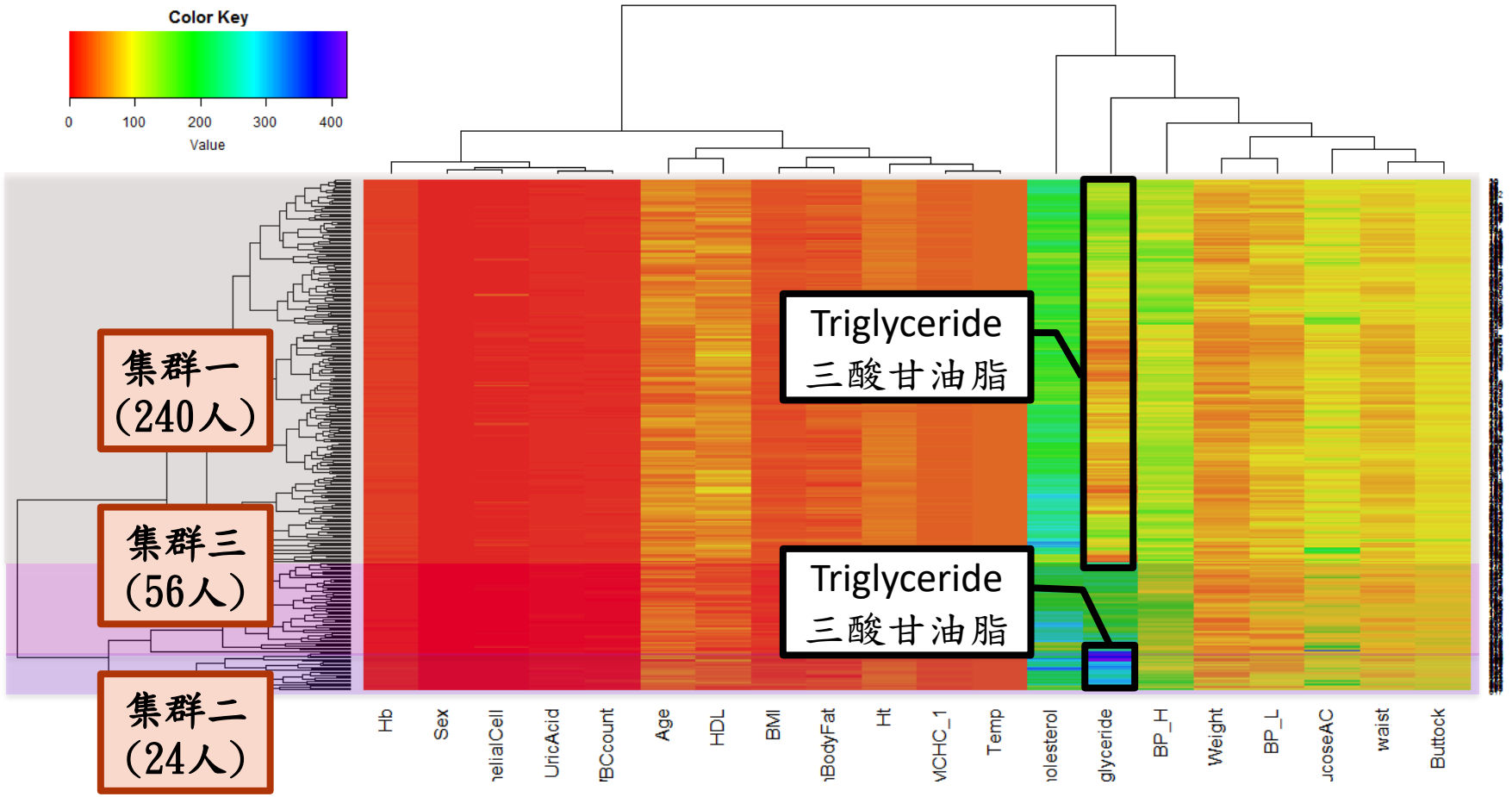
階層式集群分析

Heatmap

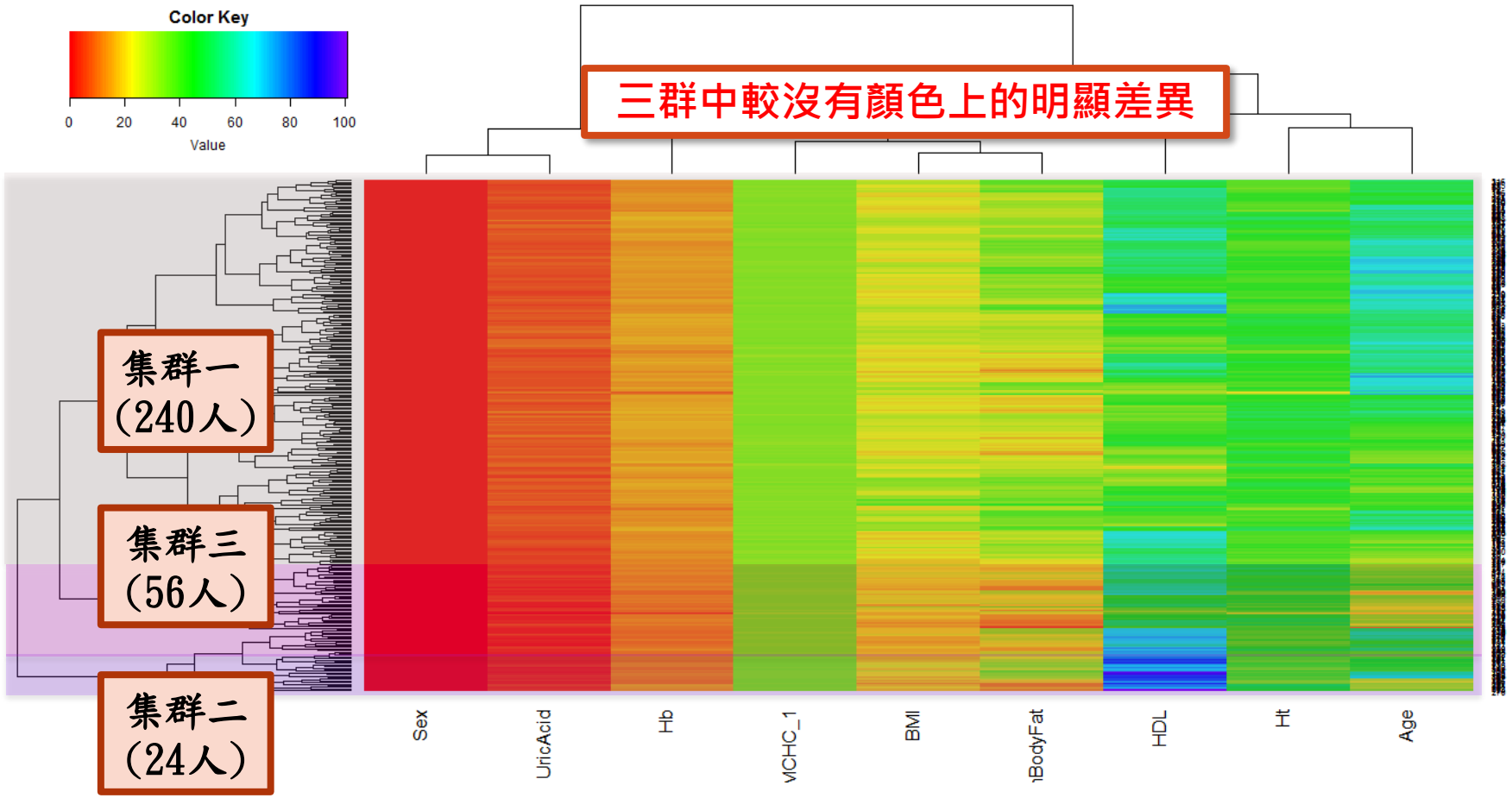


階層式集群分析

Heatmap–完整連結

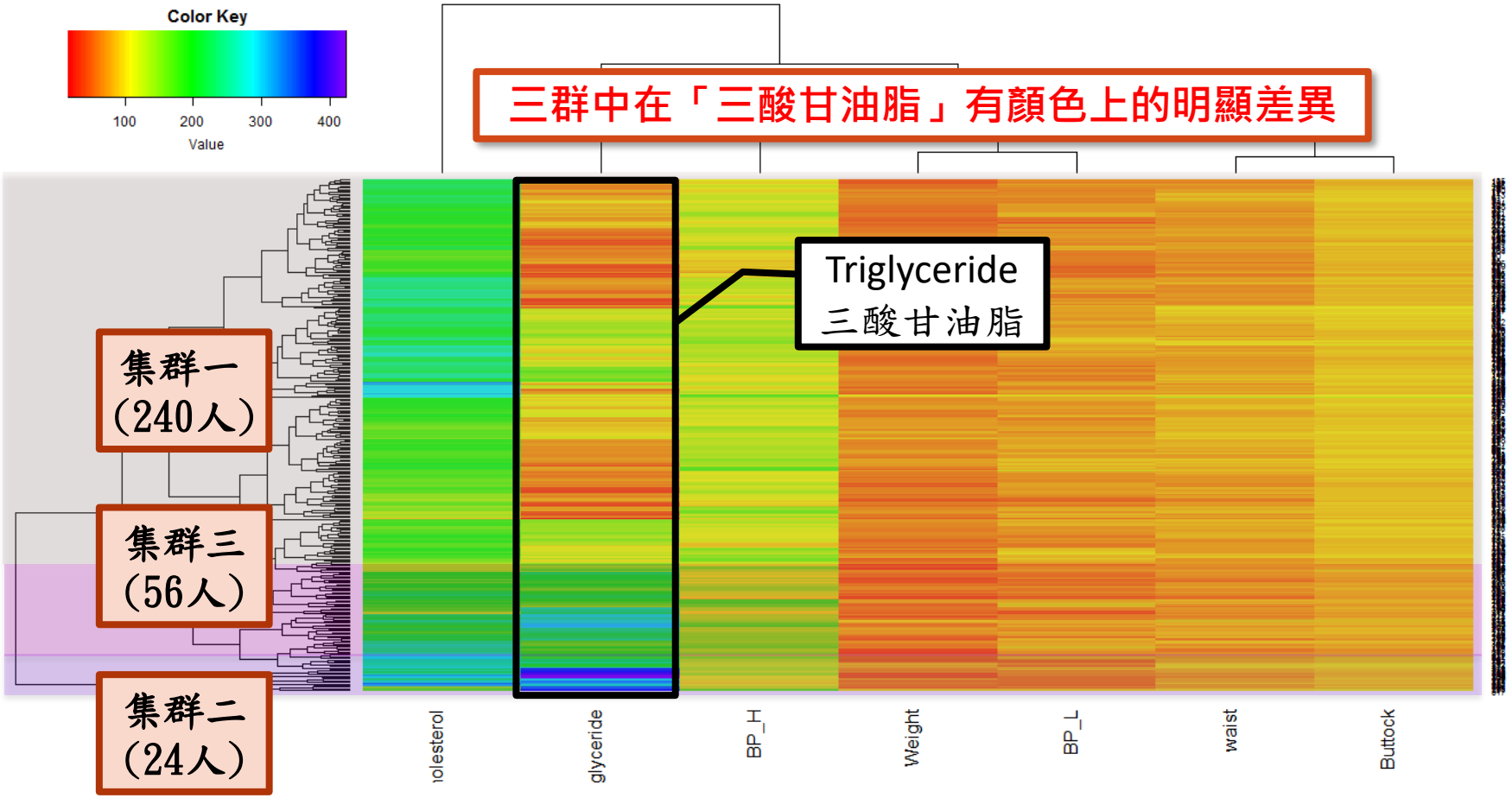


Heatmap–P.33左邊變數群，刪除EpithelialCell、Temp



階層式集群分析

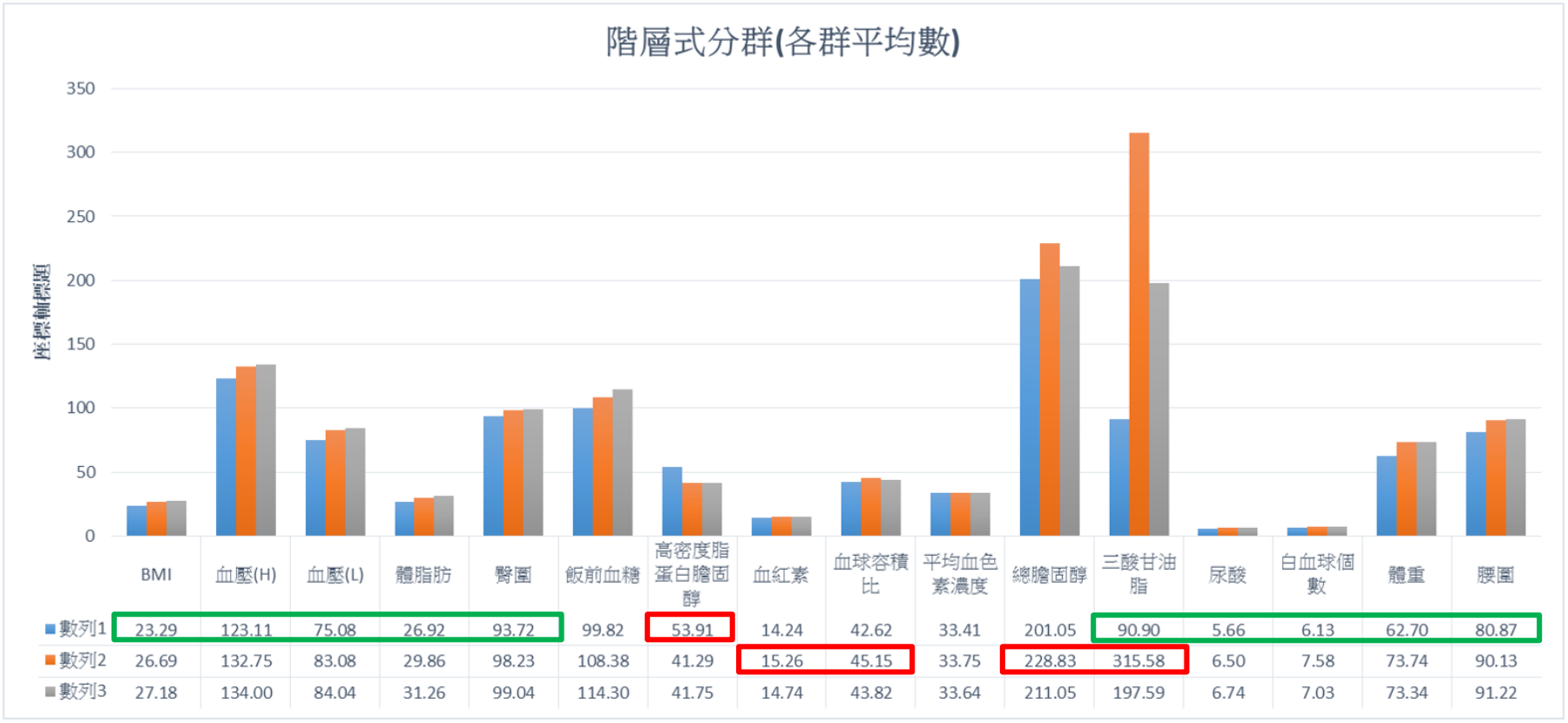
Heatmap-P.33右邊變數群，刪除GlucoseAC、WBCcount



階層式集群分析

分群結果

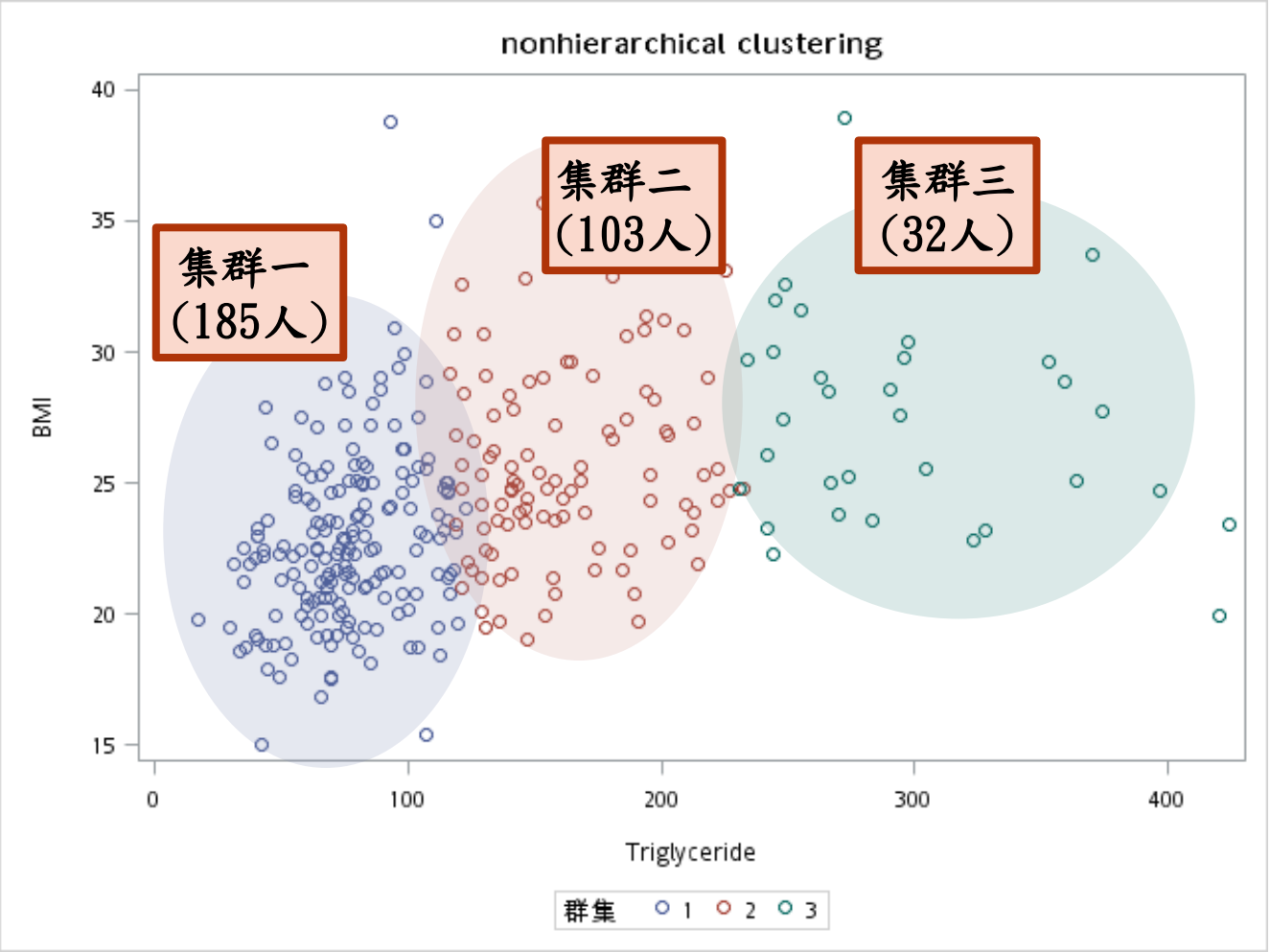
不顯著變數：年齡(Age)、上皮細胞(EpithelialCell)、性別(Sex)、體溫(Temp)



	BMI	血壓(H)	血壓(L)	體脂肪	臀圍	飯前血糖	高密度脂蛋白膽固醇	血紅素	血球容積比	平均血色素濃度	總膽固醇	三酸甘油脂	尿酸	白血球個數	體重	腰圍
數列1	23.29	123.11	75.08	26.92	93.72	99.82	53.91	14.24	42.62	33.41	201.05	90.90	5.66	6.13	62.70	80.87
數列2	26.69	132.75	83.08	29.86	98.23	108.38	41.29	15.26	45.15	33.75	228.83	315.58	6.50	7.58	73.74	90.13
數列3	27.18	134.00	84.04	31.26	99.04	114.30	41.75	14.74	43.82	33.64	211.05	197.59	6.74	7.03	73.34	91.22

非階層式集群分析

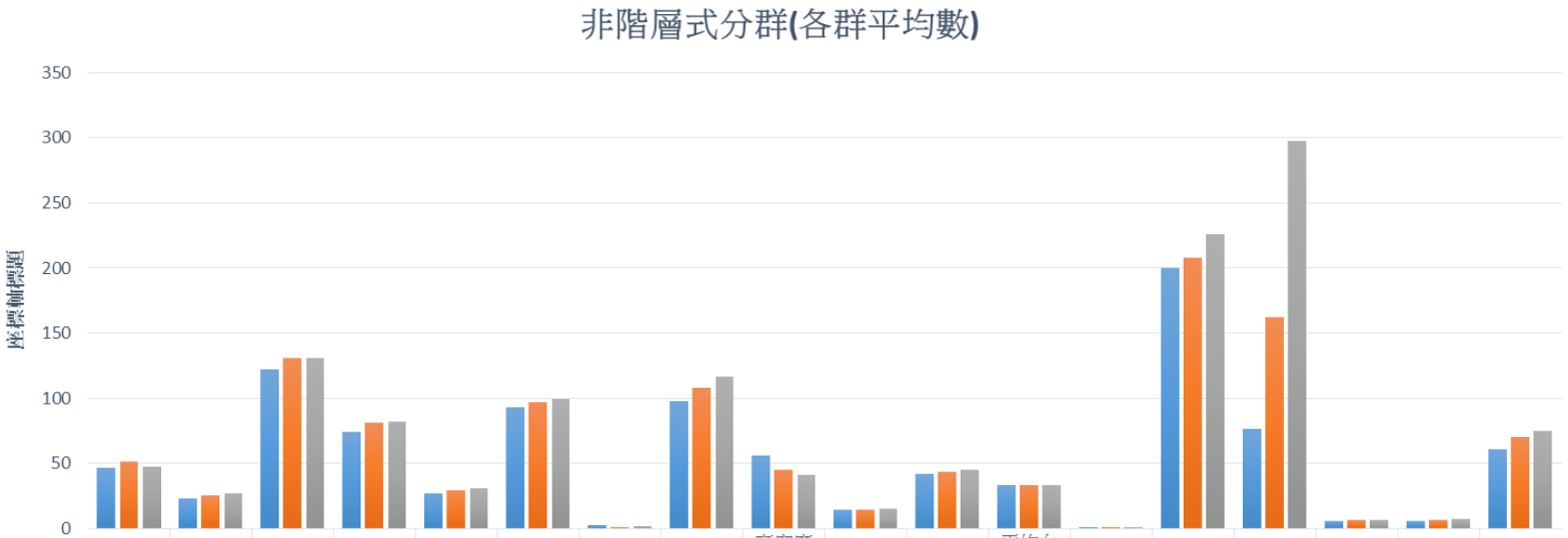
利用非階層式分成三群



非階層式集群分析

分群結果

不顯著變數：體溫(Temp)



	年齡	BMI	血壓(H)	血壓(L)	體脂肪	臀圍	上皮細胞	飯前血糖	高密度脂蛋白膽固醇	血紅素	血球容積比	平均血色素濃度	性別	總膽固醇	三酸甘油脂	尿酸	白血球個數	體重
數列1	46.72	22.85	122.21	74.19	26.65	93.13	2.38	98.01	55.82	14.09	42.22	33.36	0.46	199.59	76.78	5.51	6.00	61.05
數列2	51.00	25.74	130.45	81.33	29.10	96.94	1.25	107.78	44.90	14.73	43.82	33.61	0.68	207.78	162.42	6.46	6.73	70.17
數列3	47.69	27.33	131.03	81.78	31.27	99.48	1.47	116.44	41.13	15.19	45.06	33.68	0.63	226.19	297.56	6.47	7.61	75.11

判别分析

Fisher Linear Discriminant Function

1. 選擇變數

- 分組變數：
  a. 脂肪肝檢查正常(0)、異常(1)
  b. 脂肪肝檢查正常(0)、輕度(1)、中度以上(2)
- 區別變數：
  使用Anova分析，選擇組間差異顯著者為區別變數

2. 檢定

- 組間平均不相等
- 共變異數矩陣相等

3. 先驗機率

- 假設各組先驗機率相等
- 假設各組先驗機率依照樣本比例

4. 分類結果

- 原始分類結果
- 交叉驗證結果

Fisher Linear Discriminant Function

1. 選擇變數

- 區別變數：

從資料完整的46個變數選出23個Anova分析結果顯示組間差異顯著的

再做Manova分析確認仍然顯著

Multivariate Statistics and Exact F Statistics					
S=1 M=9.5 N=149.5					
統計值	值	F 值	分子自由度	分母自由度	Pr > F
Wilks' Lambda	0.61433947	9.00	21	301	<.0001
Pillai's Trace	0.38566053	9.00	21	301	<.0001
Hotelling-Lawley Trace	0.62776453	9.00	21	301	<.0001
Roy's Greatest Root	0.62776453	9.00	21	301	<.0001

變數	Pr > F	變數	Pr > F	變數	Pr > F
SGOT_AST	0.0756	waist	<.0001	Segment	0.1948
SGPT_ALT	0.0835	Buttock	<.0001	Eosinophil	0.3168
Height	0.3614	GlucoseAC	<.0001	Basophil	0.142
Weight	<.0001	T_Cholesterol	0.0043	Monocyte	0.1273
BMI	<.0001	Triglyceride	0.001	Lymphocyte	0.2349
IBW_L	0.3548	UricAcid	<.0001	Platelet	0.1071
IBW_U	0.3546	Creatinine	0.0693	Specific_Gravity	0.7347
BP_H	<.0001	HDL	<.0001	pH	0.0884
BP_L	<.0001	Hb	<.0001	Urobilinogen	0.1097
PulseRate	0.3639	RBC	0.0345	RBC_1	0.5052
Temp	<.0001	Ht	<.0001	WBC_1	0.4853
Sex	0.0039	MCV	0.0269	EpithelialCell	0.0495
Age	<.0001	MCH	0.0113	nAFP	0.4495
BodyFat	<.0001	MCHC_1	0.0048	nBodyFat	<.0001
IBF_L	0.4796	MCHC_2	0.0048	nCEA	0.0732
IBF_U	0.2108	WBCcount	0.0074		

Fisher Linear Discriminant Function

2. 檢定

- 組間平均不相等
- 共變異數矩陣相等

1. 相等 →使用合併樣本共變異矩陣
   →線性判別式

2. 不相等 →使用組內共變異矩陣
   →二次判別式

共變異數內矩陣的均齊性檢定

卡方	自由度	Pr > ChiSq
1326.443187	552	<.0001

因為卡方值在 0.1 層級顯著，所以會在判別函數中使用共變異數內矩陣。
參考: Morrison，D.F. (1976) 多變量統計法 第 252 頁。

Fisher Linear Discriminant Function

2. 檢定

$$\tilde{D}_t^2(x) = \left(-\frac{1}{2}\bar{X}_t' S^{-1}\bar{X}_t + \ln q_t\right) + x' S^{-1}\bar{X}_t$$

→二次判別式

在使用單個類的共變異矩陣估計時，廣義平方距離的函數為二次函數，因此稱為二是判別函數

	先驗機率均等	先驗機率依照樣本數比例
分成0、1兩組	FatLiver_01 的廣義平方距離 From FatLiver_01 | 0 | 1 0 | 7.33547 | 16.15947 1 | 12.66685 | 13.19501	FatLiver_01 的廣義平方距離 From FatLiver_01 | 0 | 1 0 | 8.92356 | 17.36248 1 | 14.25494 | 14.39802
分成0、1、2三組	FatLiver_012 的廣義平方距離 From FatLiver_012 | 0 | 1 | 2 0 | 13.42729 | 17.41037 | 28.12691 1 | 17.54369 | 14.77773 | 15.95590 2 | 28.03088 | 17.60061 | 12.25429	FatLiver_012 的廣義平方距離 From FatLiver_012 | 0 | 1 | 2 0 | 15.01538 | 19.12571 | 32.30446 1 | 19.13178 | 16.49307 | 20.13344 2 | 29.61897 | 19.31595 | 16.43183

Fisher Linear Discriminant Function

3. 先驗機率

- 假設各組先驗機率相等
- 假設各組先驗機率依照樣本比例

4. 分類結果

- **原始組別**

✓ 原始資料做出的判別式，來分每個資料，比較分類結果是否正確。

✓ 判別式使用資料中包含觀察對象

- **交叉驗證(Cross- Validation)**

✓ 保留一個觀察值，利用其它做判別分析再將此觀察值代入，比較分類結果是否正確。

✓ 判別式使用資料中不包含觀察對象

✓ 在樣本小的時候可以避免over-fitting的問題

Fisher Linear Discriminant Function

分2組:0、1

原始組別

先驗機率		0.5	0.5			
		0	1		user	錯分率
	0	139	7	146	0.95	0.048
	1	65	112	177	0.63	0.367
		204	119	323		0.208
pro		0.68	0.94		0.78	

		0.452	0.528		
		0	1	user	錯分率
0	139	7	146	0.952	0.05
1	59	118	177	0.667	0.33
	198	125	323		0.2
pro	0.702	0.944		0.796	

交叉驗證

先驗機率		0.5	0.5			
		0	1		user	錯分率
	0	121	25	146	0.83	0.171
	1	83	94	177	0.53	0.469
		204	119	323		0.32
pro		0.59	0.79		0.67	

		0.452	0.528		
		0	1	User	錯分率
0	120	26	146	0.822	0.18
1	81	96	177	0.542	0.46
	201	122	323		0.32
Pro	0.597	0.787		0.669	

Fisher Linear Discriminant Function

- 在交叉分析中的結果，錯分率會明顯上升，可以顯示原本可能因訓練樣本數較小的原因而導致over-fitting
- Fisher判別法假設各組先驗機率為相等，但實際上各組樣本數不同，先驗機率亦有差異。
- 在混淆矩陣中，若取樣的比例與原本的兩群比例不同時，會與真實的使用者正確率有差異。

Fisher Linear Discriminant Function

分3組:0、1、2

先驗機率 0.33 0.33 0.33　　0.452 0.424 0.124

原始組別

	0	1	2		user	錯分率		0	1	2		user	錯分率
0	129	15	2	146	0.884	0.116	0	131	15	0	146	0.897	0.103
1	28	104	5	137	0.759	0.241	1	28	107	2	137	0.781	0.219
2	5	4	31	40	0.775	0.225	2	9	8	23	40	0.575	0.425
	162	123	38	323		0.192		168	130	25	323		0.192
pro	0.796	0.846	0.816		0.817		pro	0.78	0.823	0.92		0.808	

0.33 0.33 0.33　　0.452 0.424 0.124

交叉驗證

	0	1	2		user	錯分率		0	1	2		User	錯分率
0	106	36	4	146	0.726	0.274	0	108	37	1	146	0.74	0.26
1	47	80	10	137	0.584	0.416	1	49	81	7	137	0.591	0.409
2	12	24	4	40	0.1	0.9	2	12	26	2	40	0.05	0.95
	165	140	18	323		0.525		169	144	10	323		0.409
pro	0.642	0.571	0.222		0.588		pro	0.639	0.563	0.2		0.591	

Fisher Linear Discriminant Function

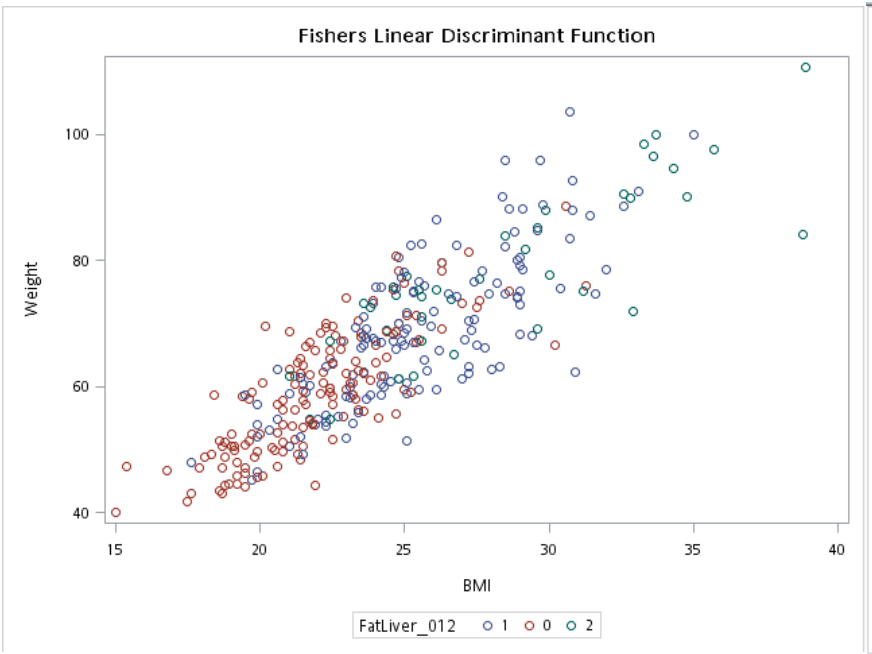
- 分為兩組的情況，先驗機率依照樣本數與均等規則比較，結果相差不遠，在各種正確率來看都是在小數點第三位以後才有差異。
- 分為三組的情況，先驗機率以按照樣本數比例做分群者，錯分率較假設先驗機率均等為佳。
- 將分為兩組與分為三組比較，這裡的先驗機率計算方式的差異影響結果較大，可從三組的樣本數比利差異看出來，分成0/1兩組，兩組的樣本數比利差不多，但分成三組0/1/2時，分組2（脂肪肝程度中度以上）者樣本數小很多。

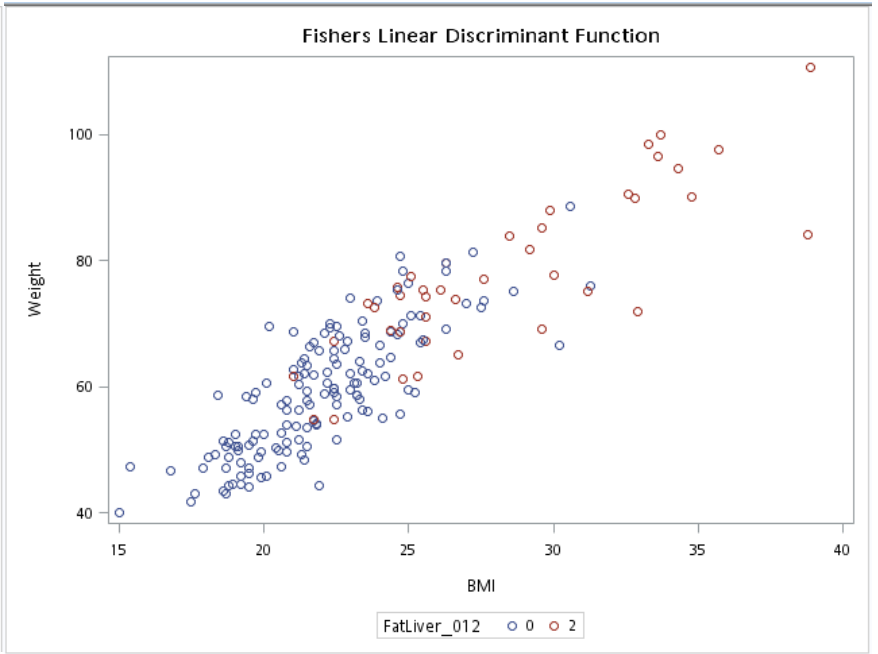
分組	0	1
樣本數比例	0.452	0.528

分組	0	1	2
樣本數比例	0.452	0.424	0.124

Fisher Linear Discriminant Function

- 考慮到脂肪肝程度0、1、2組別的分布中，1組的各項特性較分散，推測可能0與2兩組分類效果會較好，故以下嘗試0與2兩組的分析，也就是分類脂肪肝正常與脂肪肝中度以上的人。





Fisher Linear Discriminant Function

- 分組變數：脂肪肝檢查正常(0)、中度以上(2)
- 區別變數：選擇18個組間平均不相等的變數
- 共變異數矩陣相等
  → 二次判別式

FatLiver_012 的廣義平方距離			
From FatLiver_012	0	1	2
0	8.92356	13.90912	23.05994
1	13.03888	11.29113	14.42663
2	23.61398	14.16061	12.78608

變數	Pr > F				
SGOT_AST	0.2704	IBF_L	0.1148	WBCcount	0.0441
SGPT_ALT	0.1725	waist	<.0001	Segment	0.3414
AFP	0.4693	Buttock	<.0001	Eosinophil	0.6221
Height	0.1192	GlucoseAC	<.0001	Basophil	0.6837
Weight	<.0001	T_Cholesterol	0.0429	Monocyte	0.2026
BMI	<.0001	Triglyceride	0.0009	Lymphocyte	0.4107
IBW_L	0.1219	UricAcid	0.0001	Platelet	0.7313
IBW_U	0.1218	Creatinine	0.155	Specific_Gravity	0.3349
BP_H	<.0001	HDL	<.0001	pH	0.9705
BP_L	<.0001	Hb	0.0005	Urobilinogen	0.3237
PulseRate	0.5193	Ht	0.0011	RBC_1	0.8156
Temp	0.0042	MCV	0.5328	WBC_1	0.139
Sex	0.0017	MCH	0.3506	EpitheliaICell	0.0141
Age	0.0475	MCHC_1	0.0507	WBCcount	0.0441
BodyFat	<.0001	MCHC_2	0.0507		

Fisher Linear Discriminant Function

- 分類結果
- 先驗機率均等

校準資料的分類摘要: WORK.DS1
使用下列方式的交叉驗證摘要 二次判別函數

分類為 FatLiver_012 的觀測值數目和百分比

From FatLiver_012	0	2	總計
0	139 95.21	7 4.79	146 100.00
2	25 62.50	15 37.50	40 100.00
總計	164 88.17	22 11.83	186 100.00
先驗值	0.5	0.5	

FatLiver_012 的誤差計數估計值

	0	2	總計
比率	0.0479	0.6250	0.3365
先驗值	0.5000	0.5000	

- 先驗機率依樣本比例

校準資料的分類摘要: WORK.DS1
使用下列方式的交叉驗證摘要 二次判別函數

分類為 FatLiver_012 的觀測值數目和百分比

From FatLiver_012	0	2	總計
0	142 97.26	4 2.74	146 100.00
2	28 70.00	12 30.00	40 100.00
總計	170 91.40	16 8.60	186 100.00
先驗值	0.78495	0.21505	

FatLiver_012 的誤差計數估計值

	0	2	總計
比率	0.0274	0.7000	0.1720
先驗值	0.7849	0.2151	

→0與2類的分類結果較好

Nonparametric Discriminant Analysis

- 無母數統計法:不使用常態分配理論為基礎
- 嘗試最近鄰(k)=1、2、3
- 共有資料323筆，隨機選擇資料

分組	0	1	2	all
總數	146	137	40	323
Training data	140	131	38	309
Test data	6	6	2	14

分組	0	1	all
總數	146	177	323
Training data	140	163	303
Test data	6	14	20

Nonparametric Discriminant Analysis

分2組:0、1

錯分率

	k=1			k=2			k=3		
先驗機率	0.5	0.5		0.5	0.5		0.5	0.5	
組別	0	1		0	1		0	1	
train	0.21	0.20	0.20	0	0.39	0.20	0	0	0
test	0.83	0.14	0.49	0.83	0.36	0.60	0.83	0.14	0.49

錯分率

	k=1			k=2			k=3		
先驗機率	0.46	0.54		0.46	0.54		0.46	0.54	
組別	0	1		0	1		0	1	
train	0.21	0.20	0.20	0.44	0.39	0.42	0	0	0
test	0.83	0.14	0.46	1.00	0.36	0.65	0.83	0.14	0.46

- train data分類結果較test分類結果好
- 因為test data較少，錯分率的變動幅度受資料影響非常大
- 以最近鄰取3的分類結果最佳

Nonparametric Discriminant Analysis

分3組:0、1、2

錯分率

	k=1				k=2				k=3			
先驗機率	0.33	0.33	0.33		0.33	0.33	0.33		0.33	0.33	0.33	
組別	0	1	2		0	1	2		0	1	2	
train	0.24	0.47	0	0.24	0.44	0.17	0	0.20	0	0	0	0
Test	0.83	0.14		0.49	0.83	0.36		0.60	0.83	0.14		0.49

錯分率

	k=1				k=2				k=3			
先驗機率	0.45	0.42	0.12		0.45	0.42	0.12		0.45	0.42	0.12	
組別	0	1	2		0	1	2		0	1	2	
train	0.19	0.34	0.66	0.31	0.44	0.56	0.87	0.54	0	0	0	0
test	0.83	0.33	1.00	0.64	0.83	0.50	1.00	0.71	0.83	0.33	0.50	0.58

- 在先驗機率均等時，分2組及3組錯分率結果相同
- 在先驗機率依樣本比例時，分2組比分三組的錯分率結果較好

羅吉斯分析

Logistic

讀取的觀測值數目	323
使用的觀測值數目	323

最大概度估計值的分析

參數	自由度	估計值	標準誤差	Wald 卡方	Pr > ChiSq
Intercept	1	19.3796	2.5401	58.2079	<.0001
BodyFat	1	-0.1172	0.0291	16.2513	<.0001
waist	1	-0.0935	0.0205	20.8548	<.0001
GlucoseAC	1	-0.0356	0.0113	9.8648	0.0017
Hb	1	-0.3514	0.1276	7.5812	0.0059

勝算比估計值

效果	點估計值	95% Wald 信賴界限	
BodyFat	0.889	0.840	0.942
waist	0.911	0.875	0.948
GlucoseAC	0.965	0.944	0.987
Hb	0.704	0.548	0.904

Logistic

勝算比估計值			
效果	點估計值	95% Wald 信賴界限	
BodyFat	0.889	0.840	0.942
waist	0.911	0.875	0.948
GlucoseAC	0.965	0.944	0.987
Hb	0.704	0.548	0.904

- **BodyFat** Exp(-0.1172)=0.889

體脂肪增加1%，脂肪肝異常的勝算比變為原先的0.889倍

- **Waist** Exp(-0.0935)=0.911

腰圍增加1單位，脂肪肝異常的勝算比變為原先的0.911倍

- **GlucoseAC** Exp(-0.0356)=0.965

飯前血糖上升1單位，脂肪肝異常的勝算比變為原先的0.965倍

- **Hb** Exp(-0.3514)=0.704

血色素上升1單位，脂肪肝異常的勝算比變為原先的0.704倍

Logistic

- 在Logistic分析中，討論各變因對罹患脂肪肝之影響
- $\log(\frac{P_{FatLiver\_1}}{1-P_{FatLiver\_1}}) = 19.3796 - 0.1172\ X_{BodyFat} - 0.0935\ X_{waist} - 0.0356\ X_{GlucoseAC} - 0.3514\ X_{Hb}$
- 經過選模後選出4個因子，其中Hb(血色素)增加，罹患脂肪肝的的可能性降低最多，勝算比減為原先的0.704倍；GlucoseAC(飯前血糖)則影響最小，僅變為原先的0.965倍。

典型相關分析

典型相關分析

- 在典型相關分析裡有p個X變項，有q個Y變項 (p,q>1)
- 典型相關的目的在於找出這p個X變項的加權值和這q個Y變項加權值，使這p個X變項之線性組合分數與這q個Y變項之線性組合分數之相關達到最大值。

典型相關分析

A組	
Y1	身體質量指數 BMI
Y2	腰圍waist
Y3	三酸甘油酯Triglyceride
Y4	尿酸 Uric Acid
Y5	高密度脂蛋白膽固醇 HDL

B組	
X1	血清麩氨酸轉移ㄍㄇ SGOT(AST)
X2	血清丙胺酸轉胺ㄍㄇ SGPT(ALT)
X3	甲型胎兒蛋白 AFP
X4	脂肪肝

平均值和標準差

變數	平均值	標準差	標籤
Y1	24.233231	3.999553	Y1
Y2	83.399692	11.049662	Y2
Y3	140.880000	224.255946	Y3
Y4	5.910154	1.521225	Y4
Y5	50.689231	13.272502	Y5
X1	27.950769	29.261023	X1
X2	32.360000	50.787772	X2
X3	5.516308	4.020386	X3
X4	0.673846	0.683644	X4

典型相關分析

透過第一對典型因素即可解釋91.93%的解釋度，且具顯著性。

正準相關分析

	正準相關	已調整正準相關	近似標準誤差	平方正準相關	特徵值: Inv(E)*H;= CanRsq/(1-CanRsq) 特徵值	差異	比例	累計	H0 檢定: 現行列與其後所有列的正準相關皆為零 概度比	近似F值	分子自由度	分母自由度	Pr > F
1	0.591409	0.579805	0.036124	0.349764	0.5379	0.4963	0.9193	0.9193	0.62076396	8.11	20	1049	<.0001
2	0.199859	0.162052	0.053336	0.039944	0.0416	0.0363	0.0711	0.9904	0.95467534	1.24	12	838.99	0.2528
3	0.072361	.	0.055265	0.005236	0.0053	0.0049	0.0090	0.9994	0.99439503	0.30	6	636	0.9377
4	0.019256	.	0.055535	0.000371	0.0004		0.0006	1.0000	0.99962920	0.06	2	319	0.9426

第一對典型因素間的相關係數為0.59，互相的解釋量為34.97%，第二及三對的互相解釋量分別為3.99%及0.52%

(0.349764+0.039944+0.005236+0.000371)/4=9.88%

第一組解釋累積百分比為91.93%
累積至第二組為99.04%
累積至第三組為99.94%
累積至第四組為100%

只有第一組的顯著性< .05 (<.0001)

典型相關分析

A變量 與 B變量 間的相關				
	血清麩氨酸轉移酶 SGOT(AST)	血清丙胺酸轉胺酶 SGPT(ALT)	甲型胎兒蛋白 AFP	脂肪肝(0/1/2)
身體質量指數 BMI	0.0578	0.0662	0.1041	0.5547
腰圍waist	0.1141	0.1228	0.1181	0.5426
三酸甘油酯 Triglyceride	0.0351	0.0433	0.0143	0.2326
尿酸 Uric Acid	0.1383	0.1468	0.0719	0.2786
高密度脂蛋白膽固醇 HDL	0.0434	0.0212	0.0009	-0.2997

身體質量指數 BMI與腰圍waist 對脂肪肝(0/1/2) 相關係數最大，為正相關

高密度脂蛋白膽固醇 HDL與脂肪肝(0/1/2)呈現負相關

Y1~Y5分別對chi1的貢獻量0.5593~-0.2912，紅>橘>黑

A變量 與 B變量 的正準變數間的相關					
		chi1	chi2	chi3	chi4
Y1	身體質量指數 BMI	0.5593	-0.0299	-0.0184	0.0020
Y2	腰圍waist	0.5535	0.0279	-0.0021	0.0050
Y3	三酸甘油酯 Triglyceride	0.2320	-0.0124	0.0260	-0.0156
Y4	尿酸 Uric Acid	0.2920	0.0891	0.0371	0.0045
Y5	高密度脂蛋白膽固醇 HDL	-0.2912	0.1065	-0.0452	-0.0055

B變量 與 A變量 的正準變數間的相關					
		eta1	eta2	eta3	eta4
X1	血清麩氨酸轉移ㄍㄜ SGOT(AST)	0.0956	0.1873	0.0167	-0.0040
X2	血清丙胺酸轉胺ㄍㄜ SGPT(ALT)	0.1064	0.1751	0.0321	-0.0011
X3	甲型胎兒蛋白 AFP	0.1104	0.0712	-0.0347	0.0150
X4	脂肪肝	0.5837	-0.0222	0.0015	-0.0022

X1~X4分別對1的貢獻量，紅色最大

典型相關分析

A變量 的標準化正準係數					
		eta1	eta2	eta3	eta4
Y1	Y1	0.5701	-0.9867	-0.9122	-0.5108
Y2	Y2	0.3243	1.1180	0.1870	0.7223
Y3	Y3	0.2200	0.0892	0.3130	-0.9423
Y4	Y4	0.1331	0.5527	0.4778	-0.0451
Y5	Y5	-0.0109	0.8553	-0.6680	-0.4040

eta1=a11*X1+a12*X2+a13*X3+…
eta2=a21*X1+a22*X2+a23*X3+…
eta3=a31*X1+a32*X2+a33*X3+…

典型相關分析

B變量 的標準化正準係數					
		chi1	chi2	chi3	chi4
X1	X1	-0.0588	1.2623	-2.7739	-2.2649
X2	X2	0.1310	-0.3468	3.1650	2.0698
X3	X3	0.1363	0.2780	-0.4961	0.8175
X4	X4	0.9731	-0.1961	-0.0287	-0.1609

chi1=b11*Y1+b12*Y2+b13*Y3+…
chi2=b21*Y1+b22*Y2+b23*Y3+…
chi3=b31*Y1+b32*Y2+b33*Y3+…

Conclusion

- ✓ 主成分分析:
  - 前七個主成分可以解釋趨近於80%的變異量
  - 在三維圖中，健康與有脂肪肝的人有分群的趨勢
- ✓ 因素分析：
  - 探索性分析中的主成分法，有分出幾個indicator function 代表那幾個因素，但是無法有效的解釋全部因素之間的關係
- ✓ 集群分析:
  - 利用完整連結法皆有較好的表現，非階層式的群分得較為平均
- ✓ 判別分析:
  - Fisher判別法中，假設各組先驗機率以按照樣本數比例做判別分析，效果較好
  - 無母數判別分析中，以最近鄰取3的分類結果最佳
  - 分為2組較分為3組的判別結果較好
- ✓ 羅吉斯分析:
  - 血色素Hb每增加一單位，對罹患脂肪肝的勝算之降低最顯著
- ✓ 典型相關分析:
  - 身體質量指數 BMI與腰圍waist 對脂肪肝(0/1/2) 相關係數最大，且為正相關
  - 高密度脂蛋白膽固醇 HDL與脂肪肝(0/1/2)呈現負相關

分工表

統計所　張麒仙	典型相關分析、製作結果PPT
統計所　李冠薇	集群分析、製作結果PPT
生工所　凌家宜	判別分析、製作結果PPT
動科所　張馨文	主成分分析、因素分析、製作結果PPT
生工所　張少華	羅吉斯分析、製作結果PPT

參考文獻

- 王瑞蓮、藍武祥、陳季芬、劉燦榮、蕭寧馨，台灣地區老人營養健康狀況變遷調查1999-2000老年人血漿白蛋白
- 澄清綜合醫院<認識脂肪肝>
- 期刊文獻<非酒精性脂肪肝疾病的診斷與治療>

Thank you