蔡宜諠 711378912　題目：#4.7

1. 這裡的資料要用標準化後的資料還是原始資料?

摘要統計
結果

The MEANS Procedure

變數	平均值	標準差	標準誤差	變異數	最小值	最大值	N
L500	-2.8000000	6.4086432	0.6408643	41.0707071	-10.0000000	15.0000000	100
L1000	-0.5000000	7.5712108	0.7571211	57.3232323	-10.0000000	20.0000000	100
L2000	2.0000000	10.9406110	1.0940611	119.6969697	-10.0000000	45.0000000	100
L4000	21.3500000	19.6156889	1.9615689	384.7752525	-10.0000000	70.0000000	100
R500	-2.6000000	7.1237262	0.7123726	50.7474747	-10.0000000	25.0000000	100
R1000	-0.7000000	6.3968111	0.6396811	40.9191919	-10.0000000	20.0000000	100
R2000	1.6000000	9.2899424	0.9289942	86.3030303	-10.0000000	35.0000000	100
R4000	21.3500000	19.3303942	1.9330394	373.6641414	-10.0000000	75.0000000	100

從這裡可以看的出來每個變數的變異程度較大，因此我們應該使用標準化後的資料，這樣可以讓每個變數在主成分分析中有相同的權重，也比較公平。所以後續的主成分分析採用**標準化後的資料**。#

2. 根據 eigen-value-greater-than-one 應該保留多少個主成分？

相關矩陣的特徵值

	特徵值	差異	比例	累積
1	3.92900530	2.31068353	0.4911	0.4911
2	1.61832177	0.64299699	0.2023	0.6934
3	0.97532478	0.50854261	0.1219	0.8153
4	0.46678218	0.12669219	0.0583	0.8737
5	0.34008999	0.02419879	0.0425	0.9162
6	0.31589120	0.11578007	0.0395	0.9557
7	0.20011113	0.04563749	0.0250	0.9807
8	0.15447364		0.0193	1.0000

根據表 [相關矩陣的特徵值]，保留特徵值大於 1 的主成分(3.9290, 1.6183)，應該保留下**兩個主成分**。#

3. 如果保留四個主成分，能解釋的總變異量百分比是多少？

相關矩陣的特徵值				
	特徵值	差異	比例	累積
1	3.92900530	2.31068353	0.4911	0.4911
2	1.61832177	0.64299699	0.2023	0.6934
3	0.97532478	0.50854261	0.1219	0.8153
4	0.46678218	0.12669219	0.0583	0.8737
5	0.34008999	0.02419879	0.0425	0.9162
6	0.31589120	0.11578007	0.0395	0.9557
7	0.20011113	0.04563749	0.0250	0.9807
8	0.15447364		0.0193	1.0000

根據表 [相關矩陣的特徵值] 中的累積變異百分比，可以看到前四個主成分總共**解釋了 87.37% 的變異**。#

4. 為第一個主成分命名

特徵向量								
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8
L500	0.401095	-.316964	0.158157	-.327758	0.023136	0.445904	0.329255	-.546300
L1000	0.420991	-.225464	-.051961	-.481631	-.379227	-.067458	-.033121	0.622739
L2000	0.366375	0.238593	-.470293	-.282429	0.439247	-.063800	-.525517	-.186347
L4000	0.280856	0.474154	0.429502	-.161081	0.350320	-.416927	0.426944	0.083935
R500	0.343251	-.386020	0.259319	0.487600	0.497503	0.194777	-.159351	0.342530
R1000	0.411421	-.231773	-.028854	0.372316	-.351318	-.613638	-.083678	-.361365
R2000	0.311548	0.317059	-.562933	0.391417	-.110786	0.265030	0.477816	0.146588
R4000	0.254221	0.513512	0.426223	0.159098	-.395959	0.366047	-.413935	-.050821

根據表 [特徵向量]，可以看到 L500、L1000 和 R1000 這些變數的係數相對較大，且全部都是正數，其實不是太好命名，但這裡我想出最好的名字大概就是**中低頻率損失變異**。#

5. 為第四個主成分命名。

特徵向量

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8
L500	0.401095	-.316964	0.158157	-.327758	0.023136	0.445904	0.329255	-.546300
L1000	0.420991	-.225464	-.051961	-.481631	-.379227	-.067458	-.033121	0.622739
L2000	0.366375	0.238593	-.470293	-.282429	0.439247	-.063800	-.525517	-.186347
L4000	0.280856	0.474154	0.429502	-.161081	0.350320	-.416927	0.426944	0.083935
R500	0.343251	-.386020	0.259319	0.487600	0.497503	0.194777	-.159351	0.342530
R1000	0.411421	-.231773	-.028854	0.372316	-.351318	-.613638	-.083678	-.361365
R2000	0.311548	0.317059	-.562933	0.391417	-.110786	0.265030	0.477816	0.146588
R4000	0.254221	0.513512	0.426223	0.159098	-.395959	0.366047	-.413935	-.050821

根據表 [特徵向量]，可以看到 L1000 和 R500 這些變數的係數相對較大，但因為 L1000 為負值，L1000 和 R500 之間在 PRIN4 上存在反向的特性，還可以看到所有的 L 都是負數、所有的 R 都是正數，因此在這裡命名為**左右耳中低頻率變異。#**

6. 哪個 ID 的第一個主成分分數最大？

根據 Prin1 去排序後的結果輸出為下圖。

SAS 系統

觀測值	ID	L500	L1000	L2000	L4000	R500	R1000	R2000	R4000	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8
1	55	15	20	10	60	20	20	0	25	5.48987	-2.31442	1.70773	-0.02398	0.43468	-1.15662	0.34502	0.08038
2	40	10	20	15	10	25	20	10	20	5.13901	-3.22917	-0.25958	1.23567	0.05677	-0.14362	-0.73075	0.61884
3	67	5	10	40	55	0	5	30	40	4.51593	2.06094	-2.01701	-0.46434	0.77666	0.20001	0.19056	-0.10241
4	73	0	5	45	50	0	10	15	50	3.97088	2.01785	-1.32362	-0.23725	0.81993	-0.74442	-1.44449	-0.73944
5	71	0	10	40	60	-5	0	25	50	3.67591	2.97620	-1.66691	-1.01130	0.62811	0.13550	-0.25155	0.28206
6	35	-5	10	20	45	-5	10	35	60	3.58850	2.66936	-1.68959	0.96957	-1.33409	-0.26158	0.29515	0.55134
7	78	15	15	5	35	10	15	-5	0	3.38658	-2.99060	0.81985	-0.76346	0.18585	-0.96138	0.63019	-0.56423
8	98	10	10	15	55	0	0	5	75	3.29198	1.52615	1.49021	-1.13097	-0.35319	1.12289	-0.32163	-0.30686
9	60	5	10	30	20	5	5	20	10	3.19094	-0.41235	-2.22702	-0.24479	0.83306	0.28558	-0.07466	0.07965
10	14	5	15	5	60	5	5	0	50	3.05982	0.24045	1.78305	-0.75941	-0.28757	-0.27650	0.08970	0.66713
11	18	5	0	0	50	10	10	5	65	2.84260	0.45340	2.06918	1.37337	-0.20788	0.18261	-0.06345	-0.52684
12	52	5	10	20	25	0	5	15	30	2.78208	0.12202	-1.12572	-0.41599	-0.17836	0.33699	-0.05909	-0.10053

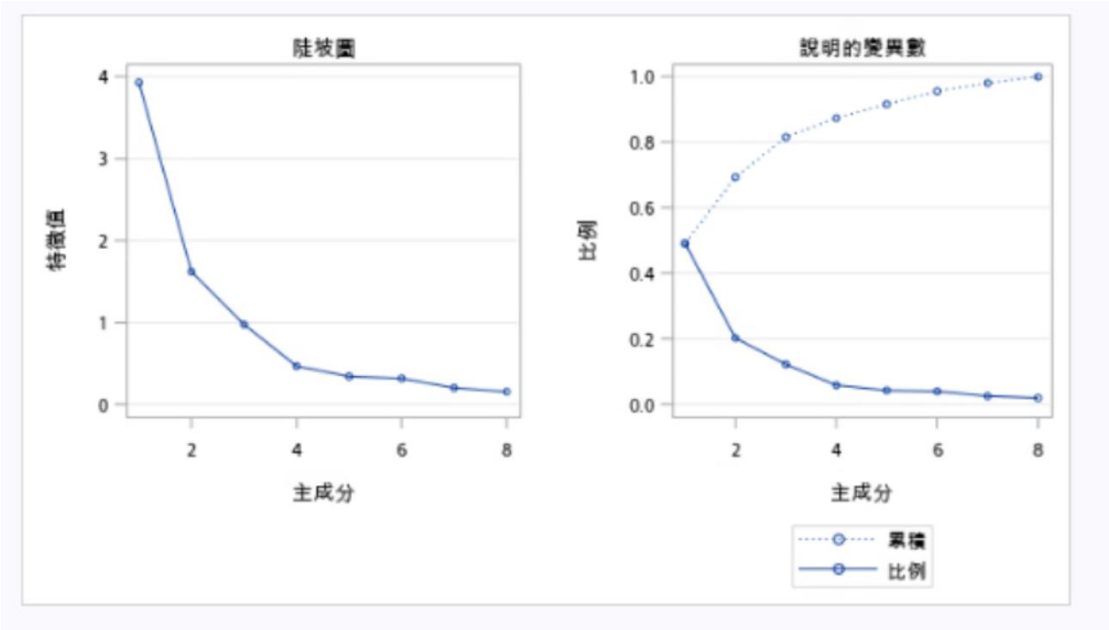
因此可以看到第一個主成分分數(5.48987)是最大的 **ID 為 55**。#

補充：

SAS 系統

觀測值	ID	L500	L1000	L2000	L4000	R500	R1000	R2000	R4000	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8
1	39	-10	-10	-10	0	-10	-10	-10	5	-3.24515	-0.23105	0.05085	-0.21346	0.05398	0.15671	-0.17597	-0.02509
2	56	-10	-10	-10	5	-10	-10	-5	-10	-3.20315	-0.33802	-0.47339	-0.16731	0.39090	-0.09097	0.51123	0.11464
3	51	-10	-10	-10	5	-10	-10	-5	0	-3.07164	-0.07237	-0.25290	-0.08501	0.18607	0.09840	0.29709	0.08835
4	80	-10	-10	-5	15	-10	-10	-5	5	-2.69526	0.41122	-0.13863	-0.25504	0.46298	-0.04863	0.16751	0.03283
5	8	-10	-10	-10	-5	-10	-5	0	5	-2.65980	-0.19178	-0.68715	0.53995	-0.42917	0.06863	0.16413	-0.17115
6	85	-10	-10	-10	5	-10	-5	-10	20	-2.65471	0.10712	0.46851	0.15995	-0.43859	-0.14516	-0.45376	-0.32558
7	41	-10	-10	-10	20	-10	-10	0	5	-2.62343	0.59368	-0.11720	0.04364	0.29191	0.01690	0.77367	0.21828
8	20	-10	-10	-5	0	-10	-5	-5	5	-2.58845	-0.13253	-0.48962	0.15915	-0.07951	-0.20945	-0.22438	-0.31381

從這個升序的表中可以看出來 ID 為 39 的的第一個主成分分數(5.48987)是最小的。



從此圖可以看出第一個主成分開始快速下降，在前兩個主成分可以解釋大約 70% 的變異，而前四個主成分可以解釋超過 85% 的變異，而後面的變異就趨於穩定，且帶來的解釋變異有限。我們可以保留 2 到 4 個主成分獲得較高的解釋力，也不會增加後續建立模型的負擔。