政 治 大 學
統 計 學 系 碩 士 班

多 變 量 分 析
期 中 報 告
Multivariate Analysis: Midterm Exam

授課教授： 洪英超 教授

研 究 生： 林健宏 統碩一 106354003

研 究 生： 曹立諭 統碩一 106354012

目錄

Question 01.

Perform a complete Principal Components Analysis for this data and interpret the result.
Note: The number of PCs must be determined by a formal statistical hypothesis test, while the relationships among objects and variables can be interpreted by using a 2D plot.

先看看資料的樣式

原始資料

Country	Agr	Min	Man	PS	Con	SI	Fin	SPS	TC
Belgium	3.3	0.9	27.6	0.9	8.2	19.1	6.2	26.6	7.2
Denmark	9.2	0.1	21.8	0.6	8.3	14.6	6.5	32.2	7.1
France	10.8	0.8	27.5	0.9	8.9	16.8	6.0	22.6	5.7
W. Germany	6.7	1.3	35.8	0.9	7.3	14.4	5.0	22.3	.6.1

(僅顯示前四資料)

資料總共有 10 個變數，26 筆樣本資料，其中 Country 為國家的名稱，其餘變數代表該國從事該職業的人口百份比(上表只顯示 4 筆國家與其產業人口百分比資料)。

變數介紹

變數名稱	變數解釋	變數型態
Country	國家名稱	類別
Agr	該國從事農業的人口百分比	數值
Min	該國從事礦業的人口百分比	數值
Man	該國從事製造業的人口百分比	數值
PS	該國從事能源業的人口百分比	數值
Con	該國從事建築業的人口百分比	數值
SI	該國從事服務業的人口百分比	數值
Fin	該國從事金融業的人口百分比	數值
SPS	該國從事社會與個人服務的人口百分比	數值
TC	該國從事交通運輸業的人口百分比	數值

9 個變數皆為數值型態，但每個變數的範圍都不大相等。因此，再進行分析前要先將資料標準化，藉此消除掉變異程度過大的變數，降低分析誤差。

資料標準化

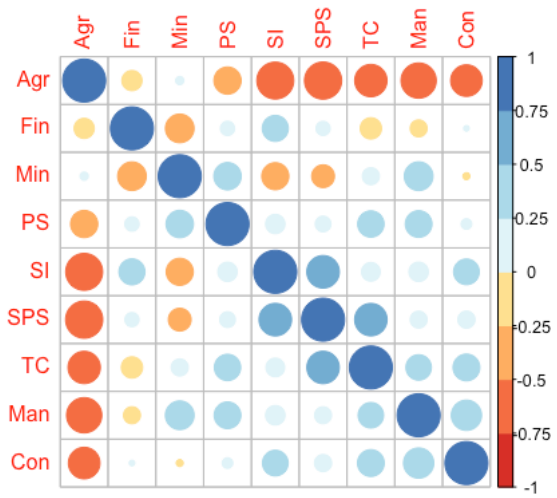
	Agr	Min	Man	PS	Con	SI	Fin	SPS	TC
Belgium	-1.0183	-0.3648	0.0845	-0.0204	0.0210	1.3425	0.7839	0.9630	0.4699
Denmark	-0.6388	-1.1895	-0.7431	-0.8179	0.0818	0.3590	0.8908	1.7830	0.3980
France	-0.5359	-0.4679	0.0703	-0.0204	0.4464	0.8398	0.7126	0.3773	-0.6081
W. Germany	-0.7996	0.0476	1.2547	-0.0204	-0.5259	0.3152	0.3563	0.3334	-0.3206

(僅顯示前面四筆)

相關係數矩陣

	Agr	Min	Man	PS	Con	SI	Fin	SPS	TC
Agr	1.0000	0.0358	-0.6711	-0.4001	-0.5383	-0.7370	-0.2198	-0.7468	-0.5649
Min	0.0358	1.0000	0.4452	0.4055	-0.0256	-0.3966	-0.4427	-0.2810	0.1566
Man	-0.6711	0.4452	1.0000	0.3853	0.4945	0.2038	-0.1558	0.1542	0.3507
PS	-0.4001	0.4055	0.3853	1.0000	0.0599	0.2019	0.1099	0.1324	0.3752
Con	-0.5383	-0.0256	0.4945	0.0599	1.0000	0.3560	0.0163	0.1582	0.3877
SI	-0.7370	-0.3966	0.2038	0.2019	0.3560	1.0000	0.3656	0.5722	0.1876
Fin	-0.2198	-0.4427	-0.1558	0.1099	0.0163	0.3656	1.0000	0.1076	-0.2459
SPS	-0.7468	-0.2810	0.1542	0.1324	0.1582	0.5722	0.1076	1.0000	0.5679
TC	-0.5649	0.1566	0.3507	0.3752	0.3877	0.1876	-0.2459	0.5679	1.0000

以圖表表示相關係數矩陣



(圖形顏色越深越大代表相關係數越大，藍色為正相關，紅色為負相關)

由相關係數表得知每兩個變數之間的相關係數都未呈現強力的正相關或負相關(相關係數 > 0.98 或 <-0.98)，而從事農業人口百分比與服務業、社會與個人服務、交通運輸業、製造業與建築業呈現中度負相關(相關係數為-0.5 ~ -0.7)，推測可能與國家主要經濟來源以及開發程度有相關。

這筆資料用於分析的變數為 9 個，過多的變數要用於統整分析資料上會有困難，不易解釋變數之間相互的影響性，同時也不易用圖表呈現。因此，我們運用除了國家名之外的 9 個變數去做組成分分析(Principal Components Analysis, PCA)，將 9 個變數透過 PCA 的方式縮減維度，同時保持數據中的對變異數貢獻最大的特徵。

Principal Components Analysis 流程

Step1. 檢定資料是否為多維常態
Step2. 檢測是否有離群值離群值(outlier)存在
Step3. 將資料的相關係數矩陣進行特徵值分解，取得特徵值（λ_i）與特徵向量（V_i）
Step4. 選擇合適的 principal components 組合
Step5. 計算 principal components (由資料內積特徵向量)、將資料投影到新的 components 上，並解釋其意義。

(一) 檢定資料是否多維常態

H_0: 資料符合多維常態
H_1: 資料不符合多維常態

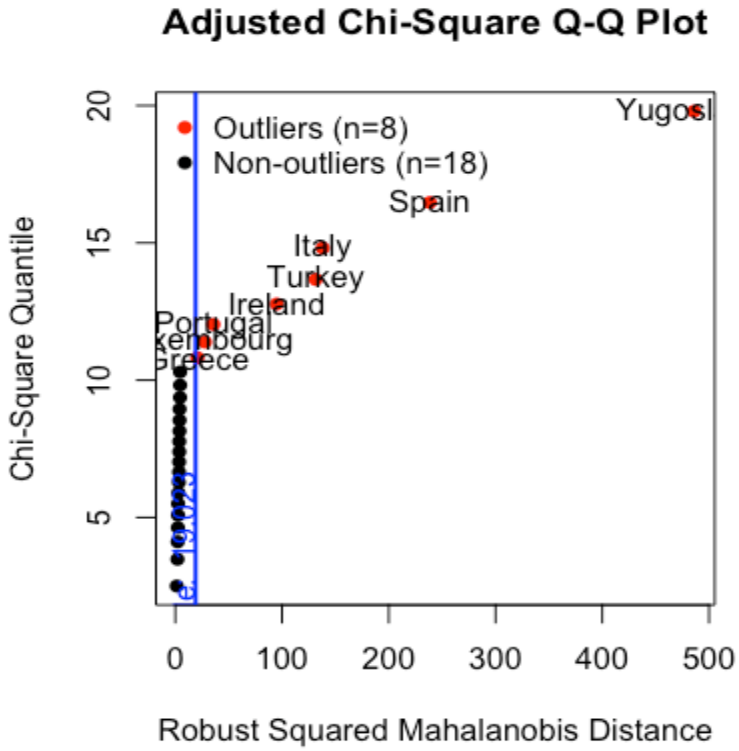
在給定顯著水準為 0.05 下，分別以不同的檢測方式檢測

Test	p-value	Result
Mardia Skewness test	0.021617	Reject H_0
Mardia Kurtosis test	0.456601	Do not reject H_0
Henze-Zirkler multinormal test	0.699563	Do not reject H_0
Royston multinormal test	0.000084	Reject H_0
Dorni-Haansen's multinormal test	0.331796	Do not reject H_0
E-statistic multinormal test	0.028	Reject H_0

上述方法有的拒絕H_0: 資料符合多維常態的假設，有的則不拒絕H_0的假設。因此我們假設資料為多元常態。

(二) 檢測離群值

由於 PCA 是根據變異數去做縮間維度的方法，因此離群值對於 PCA 方法會有很大的影響。接著使用 Robust Squared Mahalanobis Distance 方法檢測是否離群值存在，應盡量避免離群值存在，以避免資料解釋誤差產生。



圖上顯示有 8 筆資料為離群值，如 Spain, Turkey, …等 8 筆資料，但這筆資料的樣本數過少，若是任意將這 8 筆資料刪除，則資訊量會減少 30%。由於刪除離群值會造成資訊量損失過多，因此我們考慮保留這 8 筆離群值。

(三) 計算特徵向量

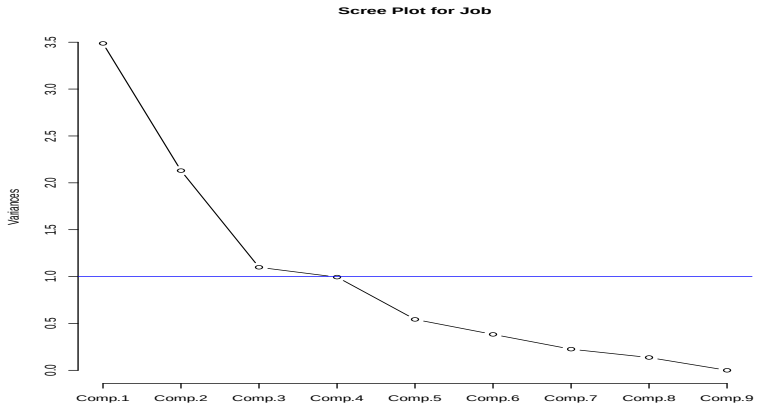
特徵向量也就是各個主成份，所對應的線性組合(linear combination)的係數、累積變異程度，接下來，我們透過R計算PCA，得到下列幾種組合以及他們解釋變異的程度

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9
Agr	-0.5238	-0.0536	0.0487	-0.0288	0.2127	-0.1533	0.0213	0.0079	0.8064
Min	-0.0013	-0.6178	-0.2011	-0.0641	-0.1637	0.1006	-0.7257	0.0884	0.0486
Man	0.3475	-0.3551	-0.1505	0.3461	-0.385	0.2882	0.4794	0.1258	0.366
PS	0.2557	-0.2611	-0.5611	-0.3933	0.2952	-0.3573	0.2556	-0.3412	0.0194
Con	0.3252	-0.0513	0.1533	0.6683	0.4716	-0.1304	-0.2207	-0.3557	0.0826
SI	0.3789	0.3502	-0.1151	0.0502	-0.2836	-0.6148	-0.2294	0.3875	0.2383
Fin	0.0744	0.4537	-0.5874	0.0516	0.2796	0.5256	-0.1875	0.1743	0.1452
SPS	0.3874	0.2215	0.3119	-0.4122	-0.2204	0.2629	-0.1913	-0.5062	0.3509
TC	0.3668	-0.2026	0.3751	-0.3144	0.5129	0.124	0.0682	0.5446	0.0721
Standard deviation	1.86739	1.45951	1.04831	0.99724	0.73703	0.61922	0.47514	0.36985	0.00675
Eigenvalue（λ_i）	3.48715	2.13017	1.09896	0.99448	0.54322	0.38343	0.22575	0.13679	0.00005
Proportion of Variance	0.38746	0.23669	0.12211	0.1105	0.06036	0.0426	0.02508	0.0152	$5.1*10^{-6}$
Cumulative Proportion	0.38746	0.62415	0.74625	0.85675	0.91711	0.95971	0.9848	0.99999	1

(四) 選取合適 components 係數

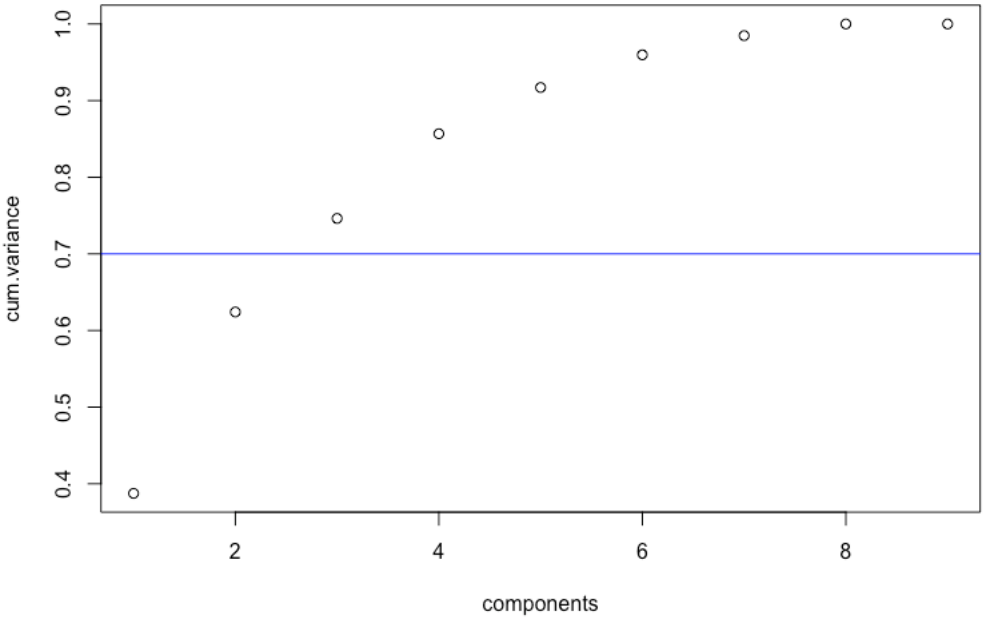
由上表去判斷要使用第幾個 principal components:

依據學者 Kaiser 提出的準則，取用$\lambda_i > 1$ 的 components 組合，則會有 components1, components2 以及 components3 作為縮減後的新維度。



（藍色線為 Eigenvalue 為 1）

依據累積總變異數比例至少 70%準則，去選取 components 係數組合，則會有 components1, components2 以及 components3 作為縮減後的新維度。



使用 Permutation test，將每一個變數的資料隨機排列（破壞 correlation structure），重新計算每個特徵值（λ_i）並記錄起來，重複上述動作多次後，繪製長條圖去看原始資料的特徵值的位置。若是原始的特徵值越大，則所對應的 p–value 則會越小。

Permutation test 的假設為

H_0: 第 i 個 component 不顯著

H_1: 第 i 個 component 顯著

經過計算後的得到 components1 ~ components9 的 p - value

	Comp.1	**Comp.2**	**Comp.3**	**Comp.4**	**Comp.5**	**Comp.6**	**Comp.7**	**Comp.8**	**Comp.9**
P - value	0.000	0.000	0.990	0.915	1.000	1.000	1.000	1.000	1.000

在給定顯著水準為 0.05 下，僅有 component1 以及 component2 有足夠的證據拒絕H_0的假設。因此在 Permutation test 下選擇 component1 與 component2 作為縮減後的二維度變數。

綜合這三中方法，前兩種方法在取捨 component 上較具有主觀意識，而 Permutation test 則是透過統計檢定方式去做 principal components 取捨。為了讓實驗更加客觀，我們選擇由 Permutation test 的取捨方式，選擇 component1 與 component2 為縮減後的二維度變數。

經過 PCA 縮減維度後，原始資料由 9 個變數構成(Agr, Min, … ,TC)降維度至 2 個變數（component1 以及 component2），而用這個兩個成分所能解釋的總變異數仍有 62%，為可接受範圍。

Component	**Agr**	**Min**	**Man**	**PS**	**Con**	**SI**	**Fin**	**SPS**	**TC**
1	-0.5238	-0.0013	0.3475	0.2557	0.3252	0.3789	0.0744	0.3874	0.3668
2	-0.0536	0.6178	0.3551	-0.2611	-0.0513	0.3502	0.4537	0.2215	-0.2026

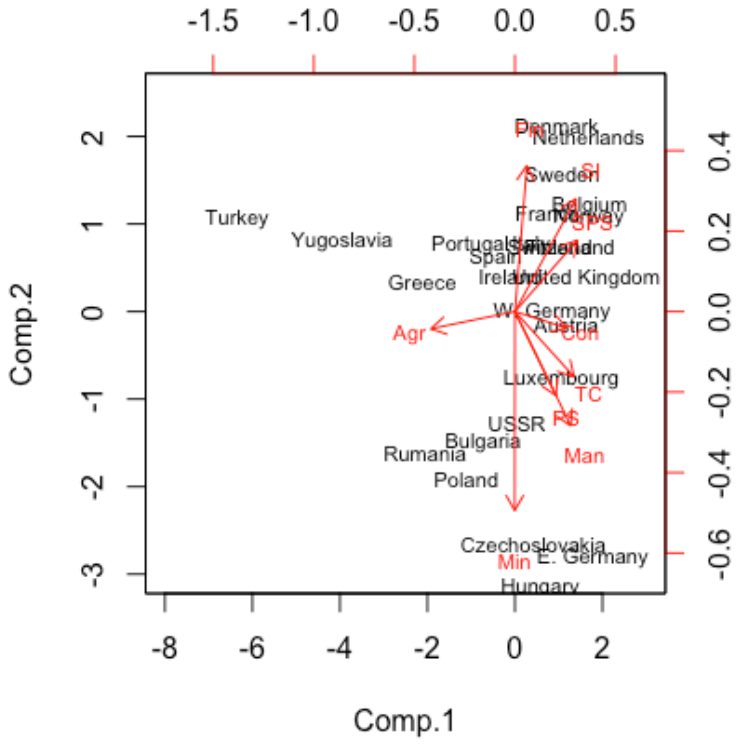
(五) 將資料投影二維平面

	Agr	Min	Man	PS	Con	SI	Fin	SPS	TC
Belgium	-1.01828	-0.36477	0.08452	-0.02045	0.02104	1.34251	0.78388	0.96301	0.46990
Denmark	-0.63878	-1.18948	-0.74313	-0.81786	0.08180	0.35895	0.89077	1.78298	0.39803
France	-0.53586	-0.46786	0.07025	-0.02045	0.44642	0.83980	0.71262	0.37732	-0.60810
W. Germany	-0.79958	0.04758	1.25465	-0.02045	-0.52588	0.31524	0.35631	0.33339	-0.32064
Ireland	0.26174	-0.26169	-0.90010	1.04277	-0.40435	0.83980	-0.42757	0.11376	-0.32064
Italy	-0.20781	-0.67404	0.08452	-1.08367	1.11487	1.12394	-0.85514	0.01126	-0.60810
Luxembourg	-0.73526	1.90317	0.54116	-0.28625	0.62872	1.21137	0.21378	-0.12052	-0.24877
Netherlands	-0.82531	-1.18948	-0.64324	0.24536	1.05410	1.10208	0.99766	1.24121	0.18243
United Kingdom	-1.05687	0.15067	0.45554	1.30858	-0.76896	0.86166	0.60572	1.21193	-0.10504
Austria	-0.41365	-0.15860	0.45554	1.30858	0.50718	0.83980	0.32068	-0.47193	0.32616

↓

	Comp.1	Comp.2
Belgium	1.71050	1.22179
Denmark	0.95290	2.12778
France	0.75463	1.12121
W. Germany	0.85255	0.01138
Ireland	-0.10350	0.41399
Italy	0.37541	0.76955
Luxembourg	1.05944	-0.75583
Netherlands	1.68822	2.00484
United Kingdom	1.63045	0.37313
Austria	1.17645	-0.14310

繪製 2 維 principal components 圖並解釋



先看水平軸 components1，數值越大表示從事農業人口百分比越低而從事建築業、服務業、製造業等的人口百分比越高，反之則相反，但從事金融業以及礦業則是不受影響。

Ex: Turkey 中從事農業人口比例有 66%其餘產業人口則不高，而 United Kingdom 從事農業人口僅有 2.7%但從事製造業、服務業性質的人口比例就很高

接著看垂直軸 components2，數值越大表示從事金融業、服務業、交通運輸的人口百分比越高而礦業、製造業、能源業等人口百分比越低，反之則相反，同樣的農業以及建築業則不受影響。

Ex: E. Germany 是屬於製造業大國，約有 41.2%的人從事該產業。

整個 2 維 principal components 圖來解釋，越靠近紅線表該產業人口所占比例越大，越右上角的國家，代表金融、服務業、交通運輸業較為發達，因此從事人口百分比較高，但礦業、製造業則越少，意味非勞力密集型的國家，可能為已開發國家居多。Ex. Sweden、France 與 United Kingdom。而右下角的國家則是顯示從事製造業、礦業以及能源業居多，可能開發中國家較多，勞力成本較低，大多公司的廠區多建在該國，但也有可能是該國本身的天然優勢使然。

Ex: Rumania、Hungary 與 E. Germany。

Question 02.

Perform a complete Canonical Correlation Analysis for these two groups of variables and interpret the result. Note: The number of canonical variates must be determined by a formal statistical hypothesis test,while the required model assumptions need to be validated.

本題使用的資料同第一題所使用，26 筆樣本資料與 10 個變數，其中 1 個變數為國家名稱，其餘 9 個為該國從事該產業的人口百分比。本題將變數分為兩類：產業屬於勞力密集型的變數集合(Agr, Min, Man, PS, Con)與產業屬於勞動較少產業(SI, Fin, SPS, TC)的變數集合。

產業屬於勞力密集型的資料

Country	Agr	Min	Man	PS	Con
Belgium	3.3	0.9	27.6	0.9	8.2
Denmark	9.2	0.1	21.8	0.6	8.3
France	10.8	0.8	27.5	0.9	8.9
W. Germany	6.7	1.3	35.8	0.9	7.3
Ireland	23.2	1	20.7	1.3	7.5

產業屬於勞動較少行業的資料

Country	SI	Fin	SPS	TC
Belgium	19.1	6.2	26.6	7.2
Denmark	14.6	6.5	32.2	7.1
France	16.8	6	22.6	5.7
W. Germany	14.4	5	22.3	6.1
Ireland	16.8	2.8	20.8	6.1

第一組資料包含 6 個變數，包含國家名稱，以及 5 個勞力密集型產業(Agr, Min, Man, PS, Con)
第二組資料包含 5 個變數，包含國家名稱，以及 4 個勞動較少產業(SI, Fin, SPS, TC)

在第一題之中，我們使用 principle components 方法縮減變異數，讓資料更為容易解釋、繪圖。而本題之中要使用的縮減變異數方法為 Canonical Correlation Analysis(CCA)，透過事先將變數分為兩組並定義每一組變數集合，接著對這兩組變數組合進行 CCA。CCA 是研究兩組變數之間相關關係的一種多元統計方法，它能解釋兩組變數之間的內在關聯同時達到降低維度的目的。

Canonical Correlation Analysis(CCA)：基本原理

Canonical Correlation Analysis(CCA)是指利用變數組合之間的相關係數來反映兩組指標（變數線性組合）之間的整體相關性的多元統計分析方法。
它的基本原理是：為了從總體上分析兩組指標之間的相關關係，分別在兩組變數中提取有代表性的兩個變數組合 U1 和 V1（分別為兩個變數組中各變數的線性組合），利用這兩個線性組合的相關關係來解釋兩組指標之間的整體相關性。

Canonical Correlation Analysis(CCA)流程:

Step1. 資料標準化，並使用標準化的資料進行以下動作
Step2. 檢定資料有 outlier 以及是否為常態
Step3. 將資料分為兩組，並定義各組
Step4. 檢測各組變數之間是否存在高度相關係數
Step5. 進行 Canonical Correlation Analysis(CCA)維度縮減
Step6. 選擇合適的成對線性組合並解釋其意義

從上表各組資料得知，每個變數的範圍皆不相同，為避免變數的變異數過大影響分析誤差，我們將資料標準化，再對各組做相關係數分析，若是任兩變數存在高度線性關係（相關係數 ＞0.98 或相關係數 ＜-0.98)，則考慮是否要移除其中一變數。

資料標準化

第一組

原始資料

Country	Agr	Min	Man	PS	Con
Belgium	3.3	0.9	27.6	0.9	8.2
Denmark	9.2	0.1	21.8	0.6	8.3
France	10.8	0.8	27.5	0.9	8.9
W. Germany	6.7	1.3	35.8	0.9	7.3
Ireland	23.2	1	20.7	1.3	7.5

標準化後資料

Country	Agr	Min	Man	PS	Con
Belgium	-1.0183	-0.3648	0.0845	-0.0204	0.0210
Denmark	-0.6388	-1.1895	-0.7431	-0.8179	0.0818
France	-0.5359	-0.4679	0.0703	-0.0204	0.4464
W. Germany	-0.7996	0.0476	1.2547	-0.0204	-0.5259
Ireland	0.2617	-0.2617	-0.9001	1.0428	-0.4043

第二組

原始資料

Country	SI	Fin	SPS	TC
Belgium	19.1	6.2	26.6	7.2
Denmark	14.6	6.5	32.2	7.1
France	16.8	6	22.6	5.7
W. Germany	14.4	5	22.3	6.1
Ireland	16.8	2.8	20.8	6.1

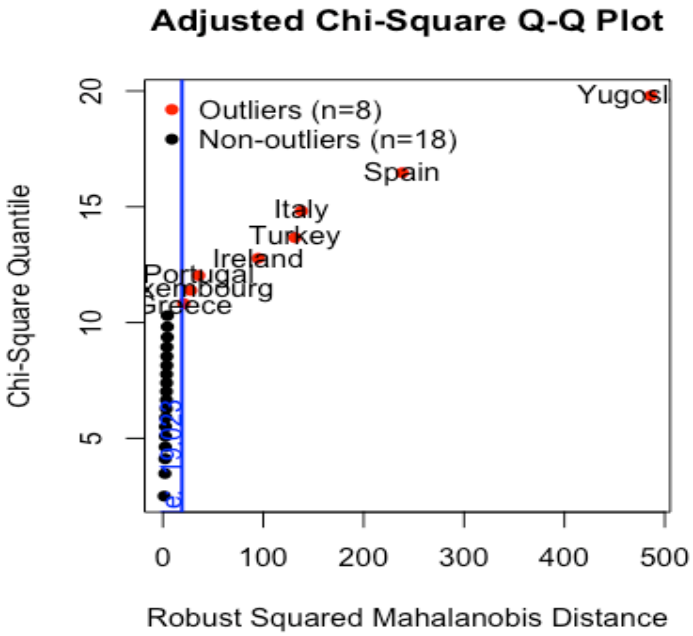
標準化後資料

Country	SI	Fin	SPS	TC
Belgium	1.3425	0.7839	0.9630	0.4699
Denmark	0.3590	0.8908	1.7830	0.3980
France	0.8398	0.7126	0.3773	-0.6081
W. Germany	0.3152	0.3563	0.3334	-0.3206
Ireland	0.8398	-0.4276	0.1138	-0.3206

檢測資料是否為常態

檢測離群值

使用 Robust Squared Mahalanobis Distance 方法檢測是否離群值存在，應盡量避免離群值存在，以避免資料解釋誤差產生。



圖上顯示有 8 筆資料為離群值，如 Spain, Turkey, …等 8 筆資料(紅點)，但這筆資料的樣本數過少，若是任意將這 8 筆資料刪除，則資訊量會減少 30%。由於刪除離群值會造成資訊量損失過多，因此我們考慮保留這 8 筆離群值。

檢定資料是否多維常態

H_0: 資料符合多維常態

H_1: 資料不符合多維常態

在給定顯著水準為 0.05 下，分別以不同的檢測方式檢測

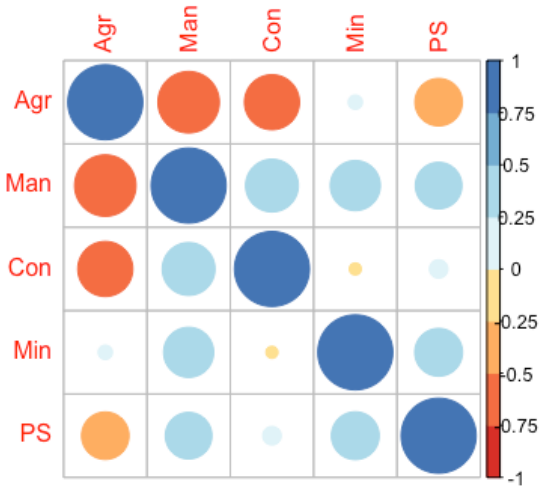
Test	p-value	Result
Mardia Skewness test	0.021617	Reject H_0
Mardia Kurtosis test	0.456601	Do not reject H_0
Henze-Zirkler multinormal test	0.699563	Do not reject H_0
Royston multinormal test	0.000084	Reject H_0
Dorni-Haansen's multinormal test	0.331796	Do not reject H_0
E-statistic multinormal test	0.028	Reject H_0

上述方法有的拒絕H_0: 資料符合多維常態的假設，有的則不拒絕H_0的假設。為了接下來分析以及縮減為度方便，我們假設資料為多元常態。

看看各組之相關係數

第一組

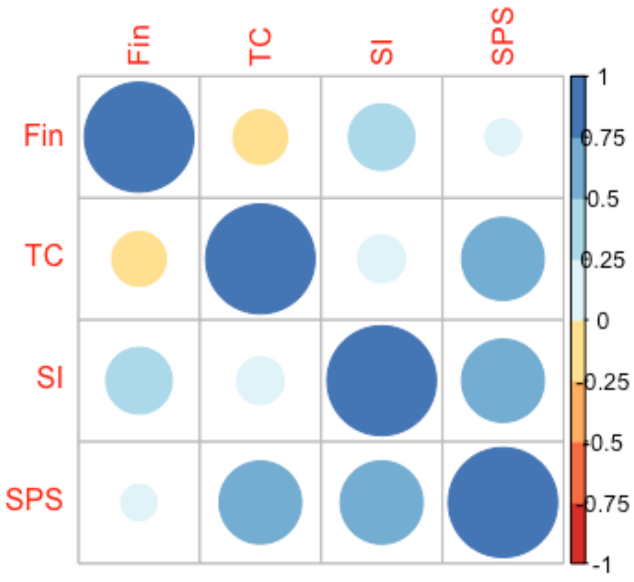
	Agr	Min	Man	PS	Con
Agr	1	0.0358	-0.6711	-0.4001	-0.5383
Min	0.0358	1	0.4452	0.4055	-0.0256
Man	-0.6711	0.4452	1	0.3853	0.4945
PS	-0.4001	0.4055	0.3853	1	0.0599
Con	-0.5383	-0.0256	0.4945	0.0599	1



(圖形顏色越深越大代表相關係數越大，藍色為正相關，紅色為負相關)

第二組

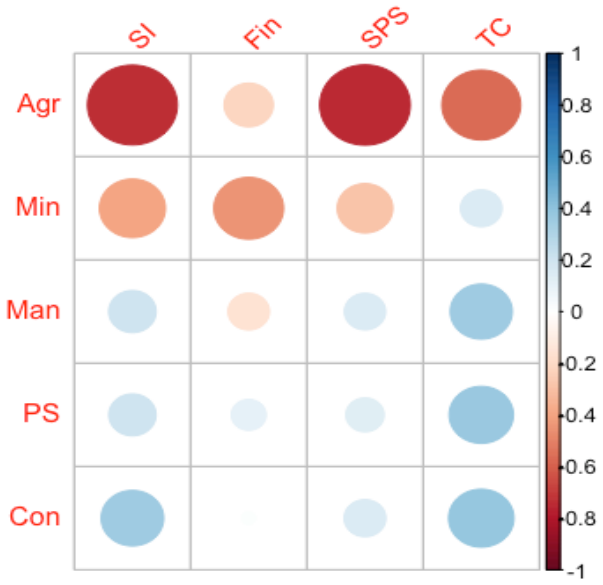
	SI	Fin	SPS	TC
SI	1	0.3656	0.5722	0.1876
Fin	0.3656	1	0.1076	-0.2459
SPS	0.5722	0.1076	1	0.5679
TC	0.1876	-0.2459	0.5679	1



(圖形顏色越深越大代表相關係數越大，藍色為正相關，紅色為負相關)

兩組之相關係數

	SI	Fin	SPS	TC
Agr	-0.7370	-0.2198	-0.7468	-0.5649
Min	-0.3966	-0.4427	-0.2810	0.1566
Man	0.2038	-0.1558	0.1542	0.3507
PS	0.2019	0.1099	0.1324	0.3752
Con	0.3560	0.0163	0.1582	0.3877



(圖形顏色越深越大代表相關係數越大，藍色為正相關，紅色為負相關)

從各組之相關係數觀察，並未出現任兩變數相關係大於 0.98 或是小於-0.98，僅有農業與服務業、社會與個人服務業呈現中度負相關。因此我們不考慮刪減變數。

進行 Canonical Correlation Analysis(CCA)維度縮減分析

透過 R 運算 CCA，結果如下

第一組變數組合之線性組合係數

	[1]	[2]	[3]	[4]
Agr	-0.2658	-0.1041	0.1771	0.0920
Min	-0.0160	-0.0016	-0.2400	-0.1238
Man	-0.1206	-0.0191	0.0564	0.2263
PS	-0.0064	-0.1490	0.1486	-0.1078
Con	-0.0272	-0.1815	0.0289	0.0439

第二組變數組合之線性組合係數

	[1]	[2]	[3]	[4]
SI	0.0786	-0.1165	0.0393	0.2195
Fin	0.0478	-0.0313	0.1616	-0.1527
SPS	0.1157	0.2453	-0.0960	-0.0786
TC	0.0238	-0.2375	-0.0301	-0.1059

第一組與第二組對應線性組合之相關係數

	[1]	[2]	[3]	[4]
Canonical Correlation(C_i)	0.9999	0.6623	0.4352	0.2152

從 Step5 中得到 4 組成對線性組合以及其相關係數，但並非每一組成對線性組合都合適，要檢測每一組成對線性組合的 Correlation 是否顯著。

使用 Wilk's test 進行檢定，而在使用 Wilk's test 檢定之前要滿足幾個條件：(1)資料樣本夠多、(2)資料服從多維常態。由於這筆資料僅有 26 筆樣本，為了檢測方便，我們認為(1)有滿足，而資料服從多維常態則在之前檢測資料是否為多維常態時，假定資料為多維常態。

Wilk's test: 檢定是否第 i 個 Canonical Correlation 為顯著

$$H_0: C_i=C_{i+1}=\dots=C_{i+k}=0$$
$$H_1: C_i \neq 0,\ C_{i+1}=\dots=C_{i+k}=0$$
$$\wedge = \prod_{i=1}^{k} 1-c_i^2$$
$$F = \frac{(1-\wedge^{\frac{1}{t}})/df_1}{\wedge^{\frac{1}{t}}/df_2} \sim F(df_1, df_2)$$

Where

$$p = \text{第一組變數個數}, q = \text{第二組變數個數}, N \text{ 為 總樣本數。}$$
$$df_1 = pq, df_2 = wt - \frac{1}{2}pq + 1, w = N - 1 - \frac{1}{2}(p+q+1), t = \sqrt{\frac{p^2q^2-4}{p^2+q^2-5}}$$
$$\text{Reject } H_0 \text{ when } F > F_{1-\alpha}(df_1, df_2)$$

檢測本題之 Canonical Correlation 是否顯著

$$H_0: C_i=C_{i+1}=\dots=C_4=0$$
$$H_1: C_i \neq 0,\ C_{i+1}=\dots=C_4=0$$

Wilks' Lambda, using F-approximation (Rao's F):

	stat	approx	df1	df2	p.value
1 to 4:	5.3764E-05	52.6824124	20	57.3325	0
2 to 4:	0.43395438	1.48132	12	47.91503	0.1644588
3 to 4:	0.77307874	0.8697839	6	38	0.5259317
4 to 4:	0.95367289	0.4857757	2	20	0.6222922

在給定顯著水準為 0.05 下，僅有 1 to 4 顯著，即只有C_1為顯著，其餘都不顯著。

經過 Wilk's test 我們選擇第一組成對線性組合係數

X' = -0.2658Agr -0.016Min – 0.1206Man – 0.0064PS – 0.0272Con

Y' = 0.0786SI + 0.0478Fin + 0.1157SPS + 0.0238TC

X 標準化後資料

Country	Agr	Min	Man	PS	Con
Belgium	-1.0183	-0.3648	0.0845	-0.0204	0.0210
Denmark	-0.6388	-1.1895	-0.7431	-0.8179	0.0818
France	-0.5359	-0.4679	0.0703	-0.0204	0.4464
W. Germany	-0.7996	0.0476	1.2547	-0.0204	-0.5259
Ireland	0.2617	-0.2617	-0.9001	1.0428	-0.4043

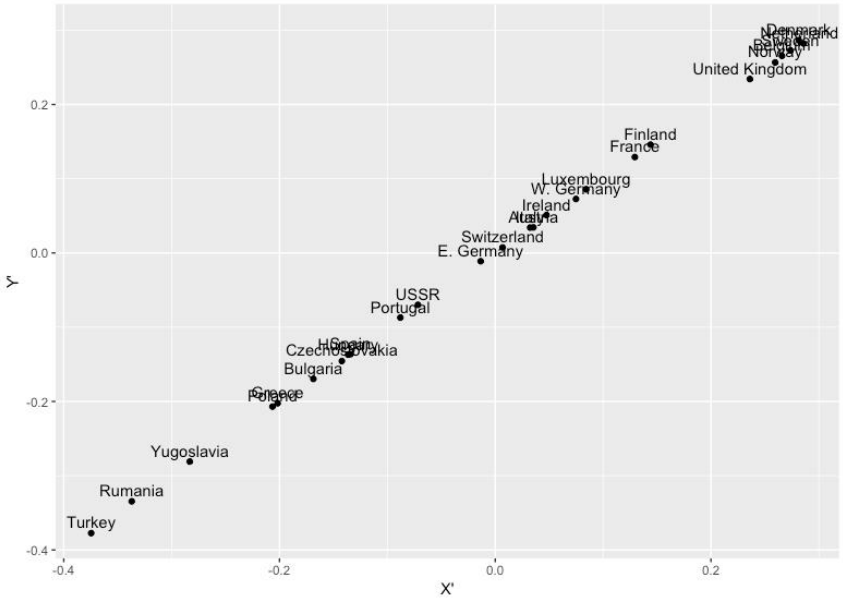
Y 標準化後資料

Country	SI	Fin	SPS	TC
Belgium	1.3425	0.7839	0.9630	0.4699
Denmark	0.3590	0.8908	1.7830	0.3980
France	0.8398	0.7126	0.3773	-0.6081
W. Germany	0.3152	0.3563	0.3334	-0.3206
Ireland	0.8398	-0.4276	0.1138	-0.3206

↓

	X’	Y’
Belgium	0.2658	0.2655
Denmark	0.2814	0.2865
France	0.1294	0.1293
W. Germany	0.0749	0.0728
Ireland	0.0475	0.0510

將資料投影到新的座標軸(X’, Y’)



解釋：

將 X’以及 Y’的係數取絕對值後，看大於 0.1 的部分，X’為 Agr 以及 Man 為最主要影響變數，而 Y’則是 SPS 為最主要影響變數。在決定 X’與 Y’兩者線性組合係數時使得 X’與 Y’的相關係數最大（本題 Corr(X’, Y’) = 0.9999），因此將資料投影到 X’與 Y’座標時，才會呈現45^{o}斜直線的形式，才顯示這筆資料的 X’與 Y’呈現強力正相關。當 X’越大時，Agr 與 Man 越小，同時 Y’與 SPS 也越大，即勞動力密集型的產業人口百分比越少時，從事勞動較少產業人口越多，反之則是。

Question 03.

Construct the decision rules for classifying the types of glass using (i) Classification Tree; (ii)LDA; (iii) QDA; (v) Nearest Neighbor; and (vi) Logistic discrimination. Compare all your resulting decision rules and explain which one you will best recommend.

（一）資料簡介

refractive	sodium	Magnesium	Aluminum	Silicon	Potassium	Calcium	Barium	Iron	Type
1.52101	13.64	4.49	1.1	71.78	0.06	8.75	0	0	1
1.51766	13.21	3.69	1.29	72.61	0.57	8.22	0	0	1
1.51743	13.3	3.6	1.14	73.09	0.58	8.17	0	0	1
1.51755	13	3.6	1.36	72.99	0.57	8.4	0	0.11	1
1.51571	12.72	3.46	1.56	73.2	0.67	8.09	0	0.24	1
1.51748	12.86	3.56	1.27	73.21	0.54	8.38	0	0.17	1

（僅顯示前六筆資料）

統計這筆資料各 Type 數

Type	1	2	3	4	5	6
個數	23	42	11	21	11	6

變數	簡介	型態
refractive	折射率	數值
sodium	鈉	數值
Magnesium	鎂	數值
Aluminum	鋁	數值
Silicon	矽	數值
Potassium	鉀	數值
Calcium	鈣	數值
Barium	鋇	數值
Iron	鐵	數值
Type	1=building_windows_float_processed 2=building_windows_non_float_processed 3=vehicle_windows_float_processed 4 = containers 5 = tableware 6 = headlamps	類別

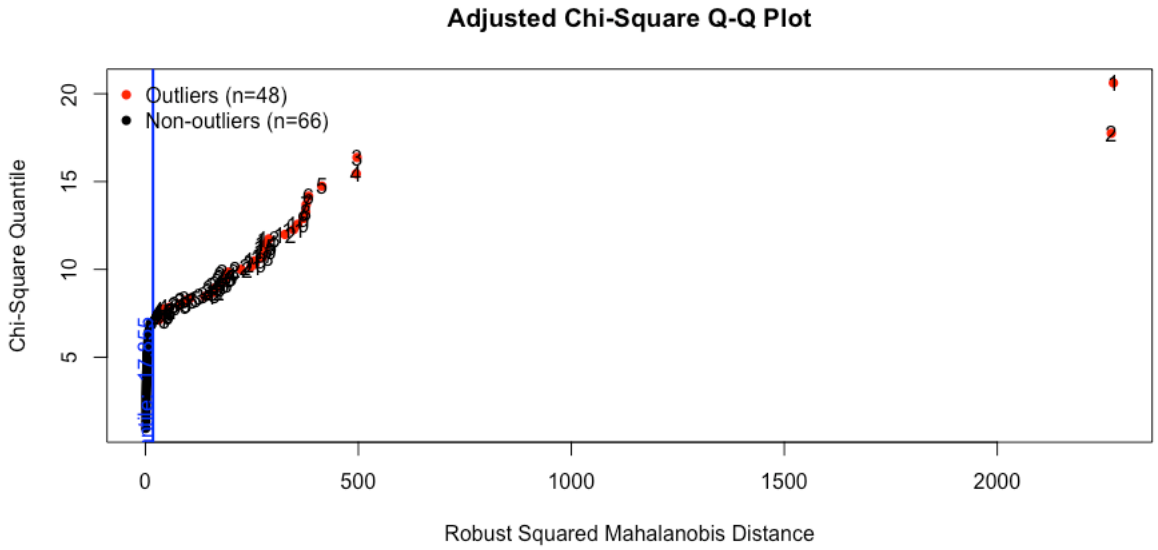
此為利用成分去歸類玻璃用途的資料，總共包含 114 樣本資料，變數共有 10 個，其中 Type 為玻璃的樣式（Type1 ~ Type6），我們要用剩下 9 個變數的資料去分類分群玻璃用途（Type）。

分類分群的方法有 Classification Tree、 LDA、 QDA、Nearest Neighbor 與 Logistic discrimination，並從中找出分類效果最好的方法(True error rate 越小越好)，藉此利用成分去分類分群未知玻璃用途（Type）。

（二）檢測離群值、共線性與資料是否為常態

檢測離群值:

由變數 Barium, Iron, Type 會造成檢測失敗，故將這三項變數拔除並檢測是否有離群值存在。使用 Robust Squared Mahalanobis Distance 方法檢測是否離群值存在，應盡量避免離群值存在，以避免資料解釋誤差產生。



由塗上很面顯看出有 48 筆離群值存在，但這筆資料僅有 114 筆資料，若是我們任意刪除這 48 筆離群值，則會造成 42%的資訊損失，因此我們不刪除任意離群值。

檢測常態:

多元常態檢定

H_0: 資料為多元常態

H_1: 資料不為多元常態

Test	Statistic	p-value	Result
Mardia Skewness	1270.60785	$33.29 * 10^{-211}$	NO
Mardia Kurtosis	36.1560802	0	NO
MVN	<NA>	<NA>	NO

個別變數常態檢定:

H_0: 變數為常態

H_1: 變數不為常態

Test	Variable	Statistic	p-value	Normality
Shapiro-Wilk	refractive	0.9412	0.00001	NO
Shapiro-Wilk	sodium	0.9228	<0.001	NO
Shapiro-Wilk	Magnesium	0.7521	<0.001	NO
Shapiro-Wilk	Aluminum	0.9473	0.00002	NO
Shapiro-Wilk	Silicon	0.9213	<0.001	NO
Shapiro-Wilk	Potassium	0.4168	<0.001	NO
Shapiro-Wilk	Calcium	0.8689	<0.001	NO

從上表發現，檢定多元常態與檢定個別變數常態都呈現拒絕H_0的假設。為了讓資料符合多元常態，我們進行了 Boxcox 的轉換，但檢定結果依然拒絕H_0的假設。因此我們推測資料可能違反多元常態的假設。

檢測是否有共線性存在

我們觀測相關係數矩陣中，是否有相關係數呈現高度相關(>0.98 或 < -0.98)

	refractive	sodium	Magnesium	Aluminum	Silicon	Potassium	Calcium	Barium	Iron
refractive	1	-0.2042	-0.0243	-0.385	-0.3468	-0.2974	0.7675	-0.2086	0.1951
sodium	-0.2042	1	-0.4909	0.1508	0.1977	-0.2853	-0.1684	0.5112	-0.2186
Magnesium	-0.0243	-0.4909	1	-0.4807	-0.2258	0.0116	-0.4045	-0.4854	0.1123
Aluminum	-0.385	0.1508	-0.4807	1	-0.1984	0.38	-0.2699	0.5692	-0.1218
Silicon	-0.3468	0.1977	-0.2258	-0.1984	1	-0.4919	-0.0046	0.0078	-0.1124
Potassium	-0.2974	-0.2853	0.0116	0.38	-0.4919	1	-0.3359	-0.0969	-0.0191
Calcium	0.7675	-0.1684	-0.4045	-0.2699	-0.0046	-0.3359	1	-0.2618	0.1509
Barium	-0.2086	0.5112	-0.4854	0.5692	0.0078	-0.0969	-0.2618	1	-0.1624
Iron	0.1951	-0.2186	0.1123	-0.1218	-0.1124	-0.0191	0.1509	-0.1624	1

從上表中觀測，別為有兩倆變數之相關係數超過 0.98 或小於-0.98，因此我們不考慮拿掉任意變數。

（三）運用不同方法去分類玻璃用途（Type）

Apparent error rate：

運用訓練樣本去建立模型，再用建立的模型去預測訓練樣本的資料，並比較預測的資料與真實樣本資料，則 Apparent error rate:為此預測錯誤的比率。

True error rate：

運用訓練樣本去建立模型，再用建立的模型去預測測試樣本的資料，並比較預的資料與真實樣本資料，則 True error rate:為此預測錯誤的比率。

在本題我們運用整筆資料去計算 Apparent error rate，並利用資料採取 cross validation(leave one out)方式去計算 True error rate.

若是 Apparent error rate 太大，則表示這方法不太用是這筆資料，若不是太大，則可以進一步計算 True error rate。

小常識：

交叉驗證 （K-fold Cross Validation）: 此法是將資料拆成 K 個(一般為 5 或 10)沒交集的群組，接著用 K-1 群的資料來建模，再用第 K 群的資料來做預測。重複這步驟 K 次，直到每一群都被用來做一次預測和被用來建立 K-1 次模型。交叉驗證可以測量模型的準確度，這對於檢視模型品質會很有幫助。而這題我們使用了 K 為樣本個數進行分群，也就是所謂的留一驗證 (leave-out-out cross validation)。

在比較方法好壞時，我們會看 True error rate 的高低，越低則表示分類準確度越高。

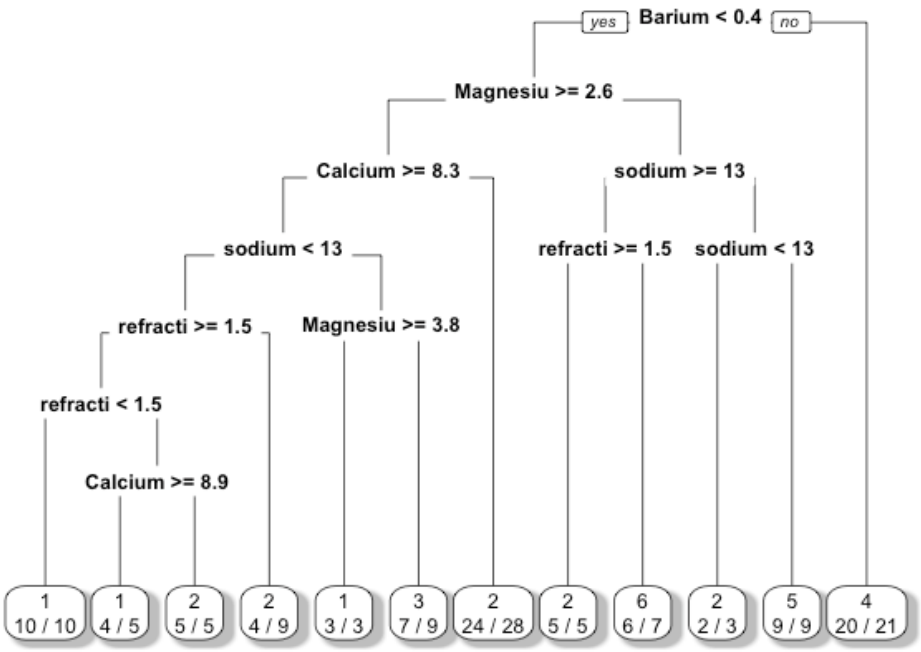
(1) Classification Tree

概念：

我們透過決策樹中的 information gaim 的方式去選擇分類樣本的變數，並設定分類準則即會產生分支，再對每一個分支重複一樣的動作，依序產生更多的分支，最終會形成一刻類似樹的樣子。

我們利用 R 的 rpart 與 prp 套件將這筆 glass 的資料使用 Classification Tree 方式去分類資料，其中 R 的參數中我們設定 minbucket 為 3 以及 minsplit 為 10 即在要分類前，該節點樣本數至少 10 筆，且末端節點至少包含 3 筆資料，這樣程式才會繼續執行。

產生以下圖示



計算結果

```
Classification tree:
rpart(formula = Type ~ ., data = glass, method = "class",
mation"),
    control = glass.control)

Variables actually used in tree construction:
[1] Barium     Calcium    Magnesium  refractive sodium

Root node error: 72/114 = 0.63158

n= 114

        CP nsplit rel error  xerror     xstd
1 0.277778      0   1.00000 1.00000 0.071533
2 0.078704      1   0.72222 1.11111 0.067842
3 0.069444      4   0.48611 0.61111 0.072192
4 0.041667      5   0.41667 0.54167 0.070352
5 0.034722      8   0.29167 0.62500 0.072485
6 0.013889     10   0.22222 0.54167 0.070352
7 0.010000     11   0.20833 0.55556 0.070772
```

Apparent error rate: $\frac{72*0.20833}{114} =$ 0.13158

True error rate: $\frac{72*0.54167}{114}$ = 0.34211

由於 **Apparent error rate 與 True error rate 差距有點大，我們認為**會產生 overfitting 的問題，會造成模型適應性不好，即帶入其他測試樣本會產生過大誤差。

(2) Linear Discriminant Analysis, LDA

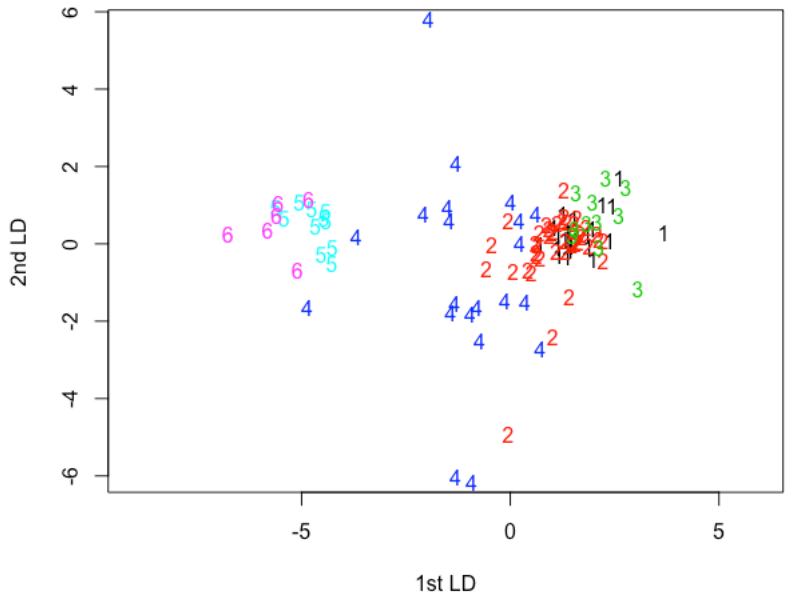
概念：

找出一條線性組合，讓這條線能夠區分不同類別的資料，並達到同群內的群內變異數小，不同群的群間變異數大的效果。

計算結果：混淆矩陣

將整筆資料帶入做訓練	cross validation(leave one out)
1 2 3 4 5 6 1 11 10 2 0 0 0 2 3 35 0 0 2 2 3 4 3 4 0 0 0 4 0 2 0 19 0 0 5 0 2 0 1 8 0 6 0 2 0 0 0 4	1 2 3 4 5 6 1 10 11 2 0 0 0 2 5 33 0 0 2 2 3 4 5 2 0 0 0 4 0 2 0 19 0 0 5 0 5 0 1 5 0 6 0 2 0 0 0 4
Apparent error rate : 0.28947	True error rate : 0.35965

（對角線上的數字為分對的個數，其餘則為分錯的個數，error rate 為分錯的比率）



上圖為我們將資料投影到 LD1 以及 LD2 的座標軸上，明顯看出 type1,2,3 混雜在一塊，而 type 5,6 混雜在一塊，說明了為何在整筆資料建立混肴矩陣時，type1,2,3 嚴重會分錯，可能在不同的 LD 可以清楚分出 type5,6。因此在混淆矩陣上 type5,6 並沒有分錯，但在 LD1, LD2 上顯示 type5, 6 混在一起。

(3) Quadratic Discriminant Analysis, QDA

概念：

類似於 LDA, 但 LDA 為畫出一直線去區分資料，QDA 則是用二次曲線的方式去將資料分類。但在做 QDA 前，資料須服從多元常態假設，但在之前的常態假設中，這筆資料並未顯示服從多元常態，且將資料進行 Boxcox 轉換後，仍不服從多元常態假設。為了練習方便，我們將這筆資料假設服從多元常態。

並且在計算 QDA 時，Barium 幾乎只有在 glass 為 containers 時才不為 0 以及 glass 為 headlamps 的個數太少，會造成 QDA 無法運算，因此我們將這個應因素去除，並計算 QDA 的 Apparent error rate 以及 true error rate.

計算結果：混淆矩陣

將整筆資料帶入做訓練	cross validation(leave one out)
1 2 3 4 5 1 19 4 0 0 0 2 10 28 4 0 0 3 0 0 11 0 0 4 0 0 0 21 0 5 0 2 0 0 9	1 2 3 4 5 1 16 7 0 0 0 2 11 26 4 0 1 3 3 5 2 1 0 4 1 1 0 19 0 5 0 7 0 1 3
Apparent error rate : 0.18519	True error rate : 0.38889

（對角線上的數字為分對的個數, 其餘則為分錯的個數，error rate 為分錯的比率）

(4) Nearest Neighbor, NN

概念：

藉由該筆樣本的最接近的 k 點去分類該筆樣本為哪一類，若該點附近皆為 A 類，則該筆樣本屬於 A 類。

在本題之中，我們分別計算 k = 1, 2, 3 的 Apparent error rate 以及 True error rate。

（樣本的應變數應為類別型態）

當 k 為 1 時，我們選擇該筆樣本資料最接近的 1 個點的類別去定義該筆樣本的類別。

混淆矩陣

將整筆資料帶入做訓練	cross validation(leave one out)
newglass.knn 1 2 3 4 5 6 1 23 0 0 0 0 0 2 0 42 0 0 0 0 3 0 0 11 0 0 0 4 0 0 0 21 0 0 5 0 0 0 0 11 0 6 0 0 0 0 0 6	newglass.knncv 1 2 3 4 5 6 1 13 6 3 1 0 0 2 8 27 3 0 2 2 3 4 1 6 0 0 0 4 0 1 0 19 0 1 5 0 0 0 1 10 0 6 0 2 0 1 0 3
Apparent error rate : 0	True error rate : 0.31579

（對角線上的數字為分對的個數，其餘則為分錯的個數，error rate 為分錯的比率）

由 Apparent error rate 為 0 可知，當 k = 1 時，會產生嚴重 overfitting 的問題，會造成模型適應性不好，即帶入其他測試樣本會產生過大誤差。因此 k = 1 並不是一個好的選擇。

當 k 為 2 時，我們選擇該筆樣本資料最接近的 1 個點的類別去定義該筆樣本的類別。

混淆矩陣

將整筆資料帶入做訓練	cross validation(leave one out)
newglass.knn 1 2 3 4 5 6 1 15 5 3 0 0 0 2 4 34 1 0 1 2 3 2 1 8 0 0 0 4 0 0 0 21 0 0 5 0 0 0 1 10 0 6 0 2 0 1 0 3	newglass.knncv 1 2 3 4 5 6 1 13 5 4 1 0 0 2 9 26 2 1 1 3 3 6 0 4 1 0 0 4 1 1 0 18 0 1 5 0 1 0 1 9 0 6 0 2 0 2 0 2
Apparent error rate : 0.20175	True error rate : 0.36842

（對角線上的數字為分對的個數，其餘則為分錯的個數，error rate 為分錯的比率）

由於 **Apparent error rate 與 True error rate 差距有點大，我們認為可能**會產生 overfitting 的問題，會造成模型適應性不好，即帶入其他測試樣本會產生過大誤差。

當 k 為 3 時，我們選擇該筆樣本資料最接近的 1 個點的類別去定義該筆樣本的類別。

混淆矩陣

將整筆資料帶入做訓練	cross validation(leave one out)
newglass.knn 1 2 3 4 5 6 1 19 3 0 1 0 0 2 3 38 0 1 0 0 3 3 0 8 0 0 0 4 0 0 0 21 0 0 5 0 0 0 1 10 0 6 0 2 0 1 0 3	newglass.kncv 1 2 3 4 5 6 1 13 6 3 1 0 0 2 5 29 2 0 3 3 3 3 2 6 0 0 0 4 0 1 1 18 0 1 5 0 2 0 1 8 0 6 0 3 0 3 0 0
Apparent error rate : 0.18421	True error rate : 0.35088

（對角線上的數字為分對的個數，其餘則為分錯的個數，error rate 為分錯的比率）

由於 **Apparent error rate 與 True error rate 差距有點大，我們認為可能**會產生 overfitting 的問題，會造成模型適應性不好，即帶入其他測試樣本會產生過大誤差。

(5) Logistic discrimination

資料的應變數為類別型態，使用全部的因變數去做 Logistic regression 模型，進而計算 **Apparent error rate 與 True error rate**

將整筆資料帶入做訓練	cross validation(leave one out)
1 2 3 4 5 6 1 16 5 2 0 0 0 2 6 35 0 0 1 0 3 1 5 5 0 0 0 4 0 1 0 20 0 0 5 0 1 0 0 10 0 6 0 0 0 0 0 6	1 2 3 4 5 6 1 12 9 2 0 0 0 2 8 29 0 1 2 2 3 4 7 0 0 0 0 4 0 1 0 19 1 0 5 1 3 0 1 6 0 6 0 0 0 1 0 5
Apparent error rate : 0.193	True error rate : 0.37719

由於 **Apparent error rate 與 True error rate 差距有點大，我們認為可能**會產生 overfitting 的問題，會造成模型適應性不好，即帶入其他測試樣本會產生過大誤差。

(6)結論

統整並比較這 5 個方法的運用於此資料的優劣

方法	Apparent error rate	True error rate
Classification Tree	0.13158	0.34211
Linear Discriminant Analysis	0.28947	0.35965
Quadratic Discriminant Analysis	0.18519	0.38889
Nearest Neighbor (k = 2)	0.20175	0.36842
Nearest Neighbor (k = 3)	0.18421	0.35088
Logistic discrimination	0.193	0.37719

根據以上表格中的 True error rate 比較，由 Classification Tree 方法做出來的 True error rate 相對其他方法所做出的 True error rate 來得低，其次是 Linear Discriminant Analysis 與 Nearest Neighbor(k = 3)兩個方法，較不建議使用 Quadratic Discriminant Analysis 方法，因為使用 QDA 的條件較為嚴厲，資料需服從多元常態，即便不服從多元常態，也是可以將資料進行 QDA 分類，但誤差可能就會較其他方法來得大。因此，對於此筆資料(glass)，我們建議使用 Classification Tree、Nearest Neighbor(k = 3)與 Linear Discriminant Analysis。

Question 04.

Based on the control of “Classification Tree” obtained in Q3, develop two decision rules by using the strategies of (i) random forest and (ii) boosting.
How do these two strategies compare with the decision rules in Q3 in terms of prediction accuracy?

- 隨機森林 random forest

隨機森林是一個集成方法，他是將幾個建立好的模型結果整合在一起，以提升預測的準確性雖然這方法提供比較好的預測，但它在推論和解適度方面就會有所限制。隨機森林由好幾個決策樹組成，而不同決策樹是由不同隨機抽取的預測變數與觀察值所組成，也正是因為由隨機建立的樹所組成的森林而得名。

流程 ：

Step 1 : 將訓練資料集用拔靴法製造出更多的樣本。
Step 2 : 生成更多的決策樹，每個決策樹是由隨機的方式抽取預測變數及觀察值所組成 (每個節點都是獨立的) ，並根據選定的變數個數找到最好的分組。
Step 3 : 生成的每棵決策樹都不進行修剪。
Step 4 : 重複 Step 1-Step 3，獲得 N 棵隨機決策樹。
Step 5 : 將 N 棵樹的預測進行投票，選取最適合的預測。

- 提升樹模型 boosting

Boosting 是一個用來提升預測準確率的方法，尤其針對決策樹。此模型的主要概念是透過連續性的建立模型來進行學習。其第一個步驟是對所有觀測值建立一個模型，而每個觀測值所被給予的權重都相等。接著提高造成模型配飾不良的觀測值權重，反之則降低權重；重複幾次過程後，最終的模型為這些小模型累積而得的成品。

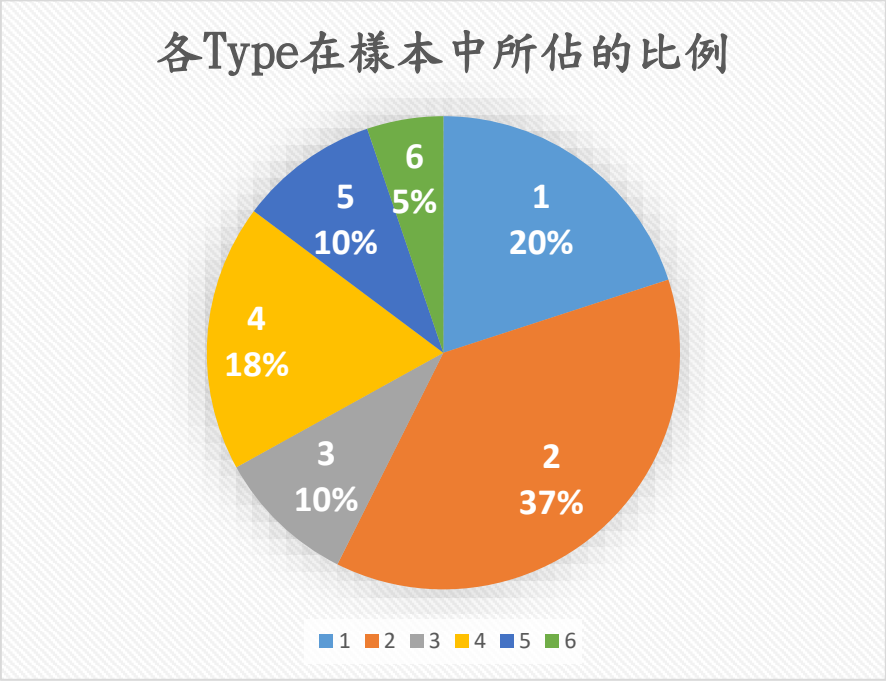
流程 ：

Step1 : 建造一棵決策樹
Step2 : 給予分類錯誤的觀察值更多的權重，並再一次進行決策樹
Step3 : 重複 Step1-Step2 過程 N 次，得到 N 次模型
Step4 : 根據觀察值的分類準確率，給予適當的投票權重，選取最適合的預測

此組資料共有 114 筆，每筆觀察值各有 9 種連續型解釋變數，以及 1 種類別型應變數
我們將使用下列兩種方法(隨機森林、提升樹模型)來進行分類

首先將此資料觀察值各 Type 個數列出

Type	1	2	3	4	5	6
個數	23	43	11	21	11	6



我們發現 Type 為 2 的個數稍微多出其他 Type，而 Type 為 6 的個數較少，可能會在抽取樣本的過程中造成我們的模型準確率下降，所以我們傾向使用交叉驗證的方法來估計真實的準確率，而不是使用切割樣本為測試及訓練資料集估計真實的準確率。

交叉驗證 （K-fold Cross Validation）:

此法是將資料拆成 K 個(一般為 5 或 10)沒交集的群組，接著用 K-1 群的資料來建模，再用第 K 群的資料來做預測。重複這步驟 K 次，直到每一群都被用來做一次預測和被用來建立 K-1 次模型。交叉驗證可以測量模型的準確度，這對於檢視模型品質會很有幫助。而這題我們使用了 K 為樣本個數進行分群，也就是所謂的留一驗證 (leave-out-out cross validation)。

OOB(Out-of-Bag)估計誤差 ：

每棵樹皆使用(boostrap)自助法來產生更多的樣本，並利用這些樣本建立決策樹模型，而沒有被抽取到的樣本就可以當作預測樣本，每個樣本都會帶入沒有使用它建立的模型進行預測，最後將得到投票出的預測值，而最後得到的預測錯誤率及為 OOB 誤差。

根據第三題的 contorl 準則來建立下列兩個模型

- Minisplit = 10　每一個 node 最少需要 10 個 data
- Minbucket = 3　在末端的 node 上最少要 3 個 data
- Xval = 114　　將資料拆成 114 群做交叉驗證

(a) Random Forest 隨機森林

我們先將資料帶入隨機森林模型，參數設定如下

- nodesize = 3　每棵決策樹終端節點的最小尺寸
- cv.fold = 114　交叉驗證的分群個數
- ntree = 500　使用 500 棵樹來建立模型
- step = 0.9　根據 step 的值作為下一次預測變數的數量

使用 R 軟體的 randomforest 套件計算出的真實錯誤率(true error rate)結果如下

	9	8	7	6	5
Error	0.2368421	0.2543860	0.2719298	0.3070175	0.3070175
	4	3	2	1	
Error	0.2894737	0.3508772	0.4473684	0.5087719	

從上圖我們得出每個節點隨機選取 7 個變數做分割所計算出的真實錯誤率會較小，
因此我們接下來選取 mtry = 9 建立出我們的隨機森林模型。

由上述參數設定所建構的隨機森林之混淆矩陣如下

	樣本觀察值					
	1	**2**	**3**	**4**	**5**	**6**
1	12	5	2	0	0	0
2	5	23	1	0	1	2
3	7	3	1	0	0	0
4	1	0	0	14	0	0
5	0	1	0	1	7	0
6	0	2	0	0	0	4

其中 OOB 估計錯誤率為 33.7%，由於每次採樣和隨機選取變量的不同，
估計的真實錯誤率會與上述有所不同。

(b) Boosting 提升樹模型

我們先將資料帶入 Boosting 提升樹模型，參數設定如下

- nodesize = 3　每棵決策樹終端節點的最小尺寸
- cv.fold = 114　交叉驗證的分群個數
- ntree = 500　使用 500 棵樹來建立模型
- boos = T　是否在每一次迭代中使用拔靴法產生樣本
- mfinal = 10

按造上述的變數設定所建立出的提升樹模型混淆矩陣如下

	樣本觀察值					
	1	**2**	**3**	**4**	**5**	**6**
1	14	5	2	1	0	0
2	6	29	6	0	4	1
3	3	2	3	0	0	0
4	0	0	0	20	1	0
5	0	3	0	0	6	0
6	0	3	0	0	0	5

從上表計算出 OOB 估計錯誤率為 32.4%

(c) 結論

根據 a、b 上述兩題的結果比較其估計的準確率，發現由 boosting 方法建立出的模型較優於隨機森林的模型， 接著我們探討兩個方法得到的重要變數排名，發現前幾個重要變數的比重都差不多，但是在排名中間的變數重要程度卻不一樣，這就代表兩模型的分類準則會有些許誤差，進而影響到最後樣本所估計的真實錯誤率了(true error rate)

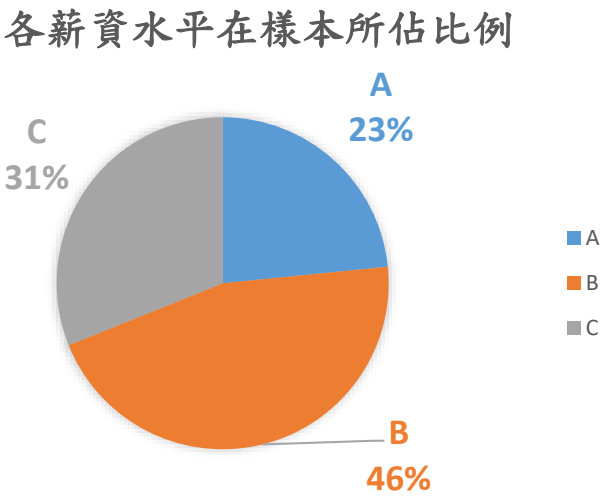
Randomforest	
refractive	6.848270
sodium	9.556605
Magnesium	11.567635
Aluminum	8.306610
Silicon	4.930122
Potassium	5.520716
Calcium	9.060250
Barium	13.601182
Iron	1.664827

Boosting	
refractive	13.159019
sodium	14.934116
Magnesium	14.445055
Aluminum	7.18747
Silicon	3.967262
Potassium	9.88821
Calcium	11.970115
Barium	24.454476
Iron	0.00000

Question 05.

Construct a suitable decision rule to classify the types of wages based on the methods introduced in our class. Also, comment on the performance of your decision rule in terms of prediction accuracy.

此題樣本共有 534 個觀察值，每個觀察值皆有 10 個類別型解釋變數，對應到 1 個類別型應變數，應變數為觀察值對應到的薪資水平 A、B、C，下圖為各薪資水平在樣本中所佔比例



從圖中我們可以發現薪資水平為 B 的觀察值所佔比例為 46%，也就是如果把樣本全部資料皆分到 B 類的正確率也有 46%，所以我們盡可能希望接下來所使用的三個方法(a)隨機森林 (b)boosting (c)羅吉斯迴歸所計算出的真實錯誤率(true error rate)能低於 50%以下，利用逐一替除交叉驗證(leave-one-out cross validation)來進行估計實際準確度，並比較三種方法的實際準確度。

(a) Random Forest 隨機森林

我們先將資料帶入隨機森林模型，參數設定如下

- cv.fold = 534　交叉驗證的分群個數
- ntree = 500　使用 500 棵樹來建立模型
- mtry = 6　每個節點分割所隨機使用的變數各數

由隨機森林所建立的模型所對應的混淆矩陣如下

	樣本觀察值		
	A	B	C
A	52	65	8
B	59	116	68
C	12	65	89

由上述混淆矩陣所得到的真實錯誤率為 51.87%

接著我們顯示隨機森林模型中各變數的重要程度

Randomforest	
Education	28.83675
South	20.15537
Gender	20.93412
Experience	20.02564
Union	17.45359
Age	25.70950
Race	25.08639
Occupation	56.53103
Sector	19.99262
Mstat	19.67012

從上述表格可以發現 Education、Age、Race、Occupation 是影響薪資水平的重要變數，其值越高對於模型的影響則越大，而錯誤率經由 500 棵樹建立的隨機森林模型有下降至 49.13%，我們猜測變數可能不是選取的很好，或許可以嘗試拿掉其中幾個變數在進行分類會有較低的錯誤率也說不定。

(b) Boosting 提升樹模型

因為資料量較大，如果使用逐一替除交叉驗證會耗時過久，這裡我們使用一般常見的 K 折交叉驗證(K-fold cross validation)，選取 K=10 作為我們切割的群數

我們先將資料帶入 Boosting 提升樹模型，參數設定如下

- nodesize = 3　每棵決策樹終端節點的最小尺寸
- cv.fold =10　交叉驗證的分群個數
- ntree = 500　使用 500 棵樹來建立模型
- boos = T　是否在每一次迭代中使用拔靴法產生樣本
- mfinal = 20　進行迭代的次數

我們計算過後得到的真實錯誤率為 47.75%。

(c) 羅吉斯迴歸

我們先將資料帶入多分類的羅吉斯迴歸模型，並使用(leave-one-out)逐一剔除交叉驗證估計出真實錯誤率，我們分別使用 glmnet 套件及 nnet 套件建構多分類的回歸模型，結果如下

	真實錯誤率
Multinom(nnet)	0.4438202
cv.flmnet(glmnet)	0.3838951

(d) 結論

	真實錯誤率
RandomForest	51.87%
Boosting	47.75%
Multinom(nnet)	44.38%
cv.flmnet(glmnet)	38.38%

從上述表格中我們發現多分類的羅吉斯迴歸模型有著較低的真實錯誤率，而隨機森林及 Boosting 模型卻有較高的錯誤率，我們猜測這可能與原始資料有關，接著我們進一步的分析資料，發現在這 534 筆樣本中，有很多筆的數據彼此間有著相同的解釋變數，而某些變數對應出的類別型的應變數卻不相同，這些數據高達 267 筆，這也導致分類器在這邊沒有強大的作用，導致就算在訓練資料集分得很好，拿來估計的測試資料有著相同解釋變數卻有不同的應變數，解釋的效果就差了許多，而迴歸模型運用了統計方法，能稍微的降低真實錯誤率。