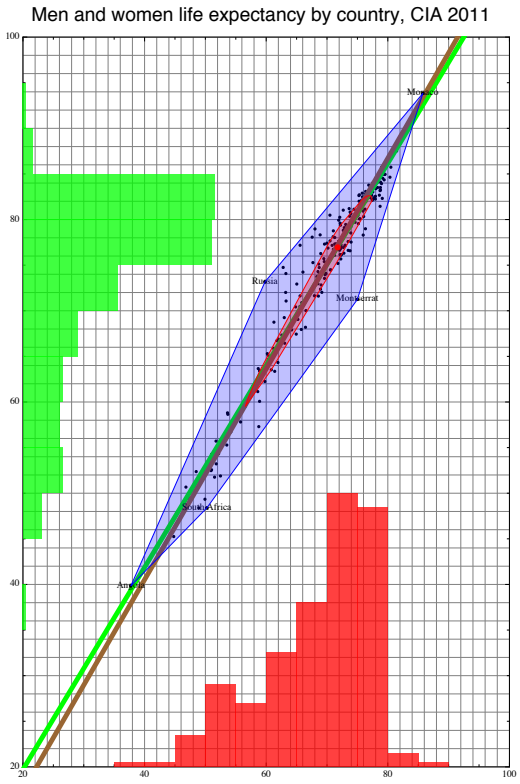


統計分析

戴忠淵

February 6, 2020

序

我打研究室走過
那獨坐電腦前的容顏如苦瓜的糾結
靈感不來，長壽的煙霧不散
研究室如小小的寂寞的城
恰如商管的電梯向晚

鍵盤不響，彈菸的手指不歇
看動作片要把小小的窗扉緊掩
我達達的馬蹄是美麗的錯誤
我不是主人，是個過客

戴忠淵
樹德科技大學經營管理研究所

目錄

1 資料的統計量數 1

1.1 集中趨勢量 2

1.2 差異量數 3

1.3 平均數與標準差的應用 4

1.4 測量尺度種類 4

1.5 常用的機率分配 5

1.6 大數法則與中央極限定理 7

1.7 常用的統計圖形 8

1.8 SPSS操作範例 10

2 區間估計與假設檢定 21

2.1 區間估計 21

2.2 假設檢定 22

2.2.1 名詞解釋 22

2.2.2 誤差之型態 22

2.3 假設檢定的型態與方法 23

2.4 母體平均數之檢定 25

2.4.1 單一母體平均數之檢定 25

2.4.2 兩獨立母體平均數差之檢定 28

2.4.3 兩相依母體平均數差之檢定 32

3 卡方檢定 37

3.1 適合度檢定 37

3.2 獨立性檢定 40
3.3 齊一性檢定 45
4 變異數分析 47
4.1 名詞解釋及基本假設 47
4.1.1 名詞解釋 47
4.1.2 假設 48
4.2 一因子變異數分析 48
4.3 二因子變異數分析 54
4.4 多重比較 59
4.5 殘差分析 60
5 簡單線性迴歸分析 63
5.1 最小平方法 63
5.2 β_0與β_1之統計推論 66
5.3 殘差分析 72
5.4 離群值偵測 79
6 多元迴歸分析 83
6.1 參數估計 83
6.2 β之統計推論 85
6.3 共線性診斷 89
7 主成份分析 99
7.1 主成份模式 99
7.2 主成份的個數選取 101
7.3 範例 102
8 因素分析 111
8.1 正交因子模式 111

8.2 因素負荷估計方法 . . . 113
8.3 共同因素個數的選取 . . . 114
8.4 因素轉軸 . . . 115
8.5 因素分數 . . . 115
8.6 範例 . . . 117
9 典型相關分析 . . . 123
9.1 典型相關模式 . . . 123
9.2 典型變量個數之選取 . . . 130
9.3 範例 . . . 130
10 區別分析 . . . 135
10.1 兩群體區別分析 . . . 135
10.1.1 資料來自常態母體，且共變異數矩陣相等 . . . 138
10.1.2 資料來自常態母體，但共變異數矩陣不相等 . . . 139
10.1.3 共變異數矩陣同質性檢定 . . . 139
10.2 多群體區別分析 . . . 140
10.3 典型區別分析 . . . 142
10.4 範例 . . . 143
A 整理資料檔案 . . . 151
A-1 分割檔案 . . . 151
A-2 選擇觀察值 . . . 152
A-3 類別化變數 . . . 155
A-4 計算 . . . 156
A-5 重新編碼 . . . 157
B Iris資料檔 . . . 159
C 機率分配表 . . . 163

索引 167
參考文獻 170

第 1 章

資料的統計量數

『統計學』是數學的一分支，主要用來搜集，分析，以及呈現資料。統計學已被廣泛的應用在各門學科之上，從自然科學到人文科學，近年來也被用來工商業及政府的情報決策上。給定一組資料，統計學可以用簡潔的方式來描述這份數據。這個方法稱作爲『描述統計學』。另外，研究者以資料的各種型態建立出數學模型，並以此來推論研究中的母體各種特徵或對未來觀察的預測，則稱爲『推論統計學』。這兩種研究方法都可以被稱作爲『應用統計學』。

研究統計學，我們必須由了解『母體』開始。所謂的母體，是由具有某些共同特性的元素或個體所組成的一個群體，是研究人員所要研究觀察的對象的全體集合。母體可能是一個國家的人民，一個學校的學生，或是某工廠生產線所生產的商品。而統計學中『母體參數』係指欲研究問題中，描述母體資料特性的統計測量數之總集合體。欲了解母體參數的特性，就必須由樣本所計算出來的統計量數，來對未知的母體做估計或利用各種統計方法，進行其推論。

本章將介紹一些統計學上常用的量化數值來解釋資料的特徵，一般在統計學上常用以描述母體的名詞有：

1 母體參數（Population Parameter）：數量化的特徵數值，用來描述母體的特性，通常唯一未知數，例如μ或σ^2等。

2 隨機樣本（Random Sample）：假設X_1, X_2, ..., X_n爲由母體$f(x)$中抽出的n個隨機變數，若滿足

(1). $X_1, X_2, \ldots, X_n$皆爲獨立

(2). $X_1, X_2, \ldots, X_n$的機率密度函數皆爲$f(x)$

則稱$X_1, X_2, \ldots, X_n$爲抽自母體$f(x)$的一組隨機樣本。

3 統計量（Statistic）：爲隨機變數的函數，並不包含未知的母體參數，且本身亦爲一隨機變數。其主要目的是用來估計母體參數，例如樣本平均數$\overline{X} = \sum_{i=1}^{n} X_i/n$及樣本變異數

$S^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2/(n-1)$皆爲統計量。

4 抽樣分配（Sampling Distribution）：統計量的機率密度函數。

1.1 集中趨勢量

集中趨勢量（Measures of Central Tendency）：資料向某一數值集中，表示此種共同趨勢 的量數，可做資料的中心代表值或兩個以上（含）群體間之比較或估計，推論母體的母體。以下介紹三種重要的集中趨勢量數（1）平均數，（2）中位數，（3）衆數。

1 平均數

(1). 算術平均數（Arithmetic Mean）：通常使用的平均數

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{1-1}$$

其中n爲樣本數。

(2). 幾何平均數（Geometric Mean）：適用於比例（等比）或變動率的數據

$$G = \sqrt[n]{X_1 \cdot X_2 \cdots X_n} = \sqrt[n]{\prod_{i=1}^{n} X_i} \tag{1-2}$$

(3). 調和平均數（Harmonic Mean）：適用速率或物價之數據

$$H = \frac{n}{\sum_{i=1}^{n}(1/X_i)} \tag{1-3}$$

2 位數

(1). 中位數（Median）：最中間的數，即其將資料群化分爲兩部份，其符號爲Me。

(2). 四分位數（Quartile）：將資料群（處理過）化分成四部分，即有三個分割點，則

$$Q_k = \begin{cases} X_{([\frac{nk}{4}]+1)}, & \text{當} kn/4 \text{不爲整數時} \\ X_{\{(\frac{nk}{4})+(\frac{nk}{4}+1)\}/2}, & \text{當} kn/4 \text{爲整數時} \end{cases} \tag{1-4}$$

(3). 十分位數（Deciles）：與前述兩種位數的概念是一樣的，符號爲 $D_1, \ldots, D_9$。

(4). 百分位數（Percentile）：同理，其符號爲$P_1, \ldots, P_9$；即 P_r，$r = 1, 2, 3, \ldots, 99$。

3 衆數（Mode）：一資料群中出現次數最多者所對應的數值。

(1). 衆數之符號常用Mo。

(2). 無任何一個數值出現超過一次以上，則衆數不存在。

(3). 衆數可能不只一個值。

(4). 如相鄰二數皆爲衆數，則取平均值。

1.2 差異量數

差異量數（Measures of Variation）：討論資料的離散程度。雖然兩組資料具有相同的$\overline{X}$或Me等時，並不代表這兩組資料群具有相同的分布情況。以下介紹三類的差異量數：

1 非離中差異量數：不以中心點爲基準，而以兩極端點爲基準，如全距。

2 離中差異量數：以中心點爲基準

(1). 以中位數Me爲中心的四分位差

$$\frac{Q_3 - Q_1}{2} \tag{1-5}$$

(2). 以中位數Me爲中心的離中平均差

$$\frac{\sum_{i=1}^{n} |X_i - \mathrm{Me}|}{n} \tag{1-6}$$

(3). 以平均數$\overline{X}$爲中心的離均平均差

$$\frac{\sum_{i=1}^{n} \left|X_i - \overline{X}\right|}{n} \tag{1-7}$$

(4). 以平均數$\overline{X}$爲中心的樣本變異數

$$\frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1} \tag{1-8}$$

1.3 平均數與標準差的應用

1 柴比雪夫定理（Chebyshev's Theorem）：不論資料分布的型態如何，若母體平均數 及母體變異數皆存在，則

$$P(|X-\mu|<k\sigma)\geq 1-\frac{1}{k^2}，k>1 \tag{1-9}$$

2 經驗法則：當資料群呈鍾型分配時，則有下列關係

(1). 大約有68% 的資料落入區間$(\overline{X}-S,\overline{X}+S)$

(2). 大約有95% 的資料落入區間$(\overline{X}-2S,\overline{X}+2S)$

(3). 大約有99.74% 的資料落入區間$(\overline{X}-3S,\overline{X}+3S)$

3 變異係數（Coefficient of Variation）：CV $=S/\overline{X}$，可使兩資料的離散程度可以比較大小。變異係數越大，離散程度越高；反之，離散程度越小。

4 標準分數：將原始資料的數值轉化爲標準值的方法

$$Z=\frac{X-\overline{X}}{S} \tag{1-10}$$

1.4 測量尺度種類

研究者在研究過程中，蒐集到資料數據中，應以何種特性或量測準則？因不同測量尺度其資料計算與進行統計推論方式均不同，因此對於「變項」的基本概念與區辨方式，是初學者進入統計學領域必備概念。而最常見測量尺度區分方式是以1951年Stevens將測量尺度區分爲四大類：

1 名目變數（Nominal Variable）：爲標示事物與事物間不同特質或類別而指定的數字，無法比較二者間大小和形式。僅衆數有意義，如：性別、血型、郵遞區號。

2 順序變數（Ordinal Variable）：不僅能標示事物與事物間不同特質或類別而指定的數字，亦能指出大小次序、多寡、優劣、高低，但無法衡量其距離差異，中位數有意義。如：考試名次、教育程度、社經地位、大學教授層級。

3 等距變數（Interval Variable）：不僅能標示事物與事物間不同特質或類別而指定的數字，亦能指出大小次序（等級順序）、多寡、優劣、高低，尚可以數值方式表示二事物差別之大小的變數。如：溫度、智商。等距變數具有「相等單位」，但無絕對零點、無倍數關係。

4 比率變數（Ratio Variable）：不僅能標示事物與事物間不同特質或類別而指定的數字，亦能指出大小次序、多寡、優劣、高低；尚可以數值方式表示二事物差別之大小的變數，還可以表示相對比例大小。比率變數最重要條件是具有絕對零點與有倍數關係，符合四則運算中加、減、乘、除運算。如：身高、年齡、體重、薪資皆爲此一變項。

1.5 常用的機率分配

以下介紹統計學中常用的一些機率分配

1 常態分配（Normal Distribution）X有常態分配以$X \sim \mathcal{N}(\mu, \sigma)$表示之，其中$\mu$爲母體期望值，$\sigma$爲母體變異數，其機率密度函數（Probability Density Ffunction，縮寫爲p.d.f.）爲

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right], \ -\infty < x < \infty \tag{1-11}$$

當期望值$\mu = 0$ 且變異數 $\sigma = 1$時，稱爲標準常態分配。

特性：

(1). 以$x = \mu$爲中心線，左右對稱

(2). σ的大小，表示圖形的高矮胖瘦

(3). X軸(即$Y = 0$)爲水平漸近線

2 卡方分配（Chi-Square Distribution）：X有卡方分配以$X \sim \chi^2(r)$表示之，其中r稱爲自由度（degree of freedom, 通常以$d.f.$表示之），其p.d.f.爲

$$f(x) = \frac{1}{\Gamma\left(\frac{r}{2}\right) 2^{\frac{r}{2}}} x^{\frac{r}{2}-1} e^{-\frac{x}{2}}, \ 0 \le x < \infty \tag{1-12}$$

$\Gamma(r) = (r-1)!$，r爲一正整數。

特性：

(1). 定義域在 $[0, \infty)$

(2). 不對稱，爲一右偏分配

(3). 自由度r不同，其圖形有所不同

(4). 爲$\alpha = \frac{r}{2}$及$\theta = 2$之 Gamma 分配

3 t分配 假設$Z \sim \mathcal{N}(0,1)$，$X \sim \chi^2(r)$且X與Y獨立，則統計量

$$T = \frac{Z}{\sqrt{\frac{X}{r}}} \sim t(r) \tag{1-13}$$

特性：

(1). 對Y軸對稱

(2). 自由度r不同，其圖形不同

(3). 與標準常態分配相似

(4). 當自由度$r > 30$時，近似標準常態分配

4 $\mathcal{F}$分配 假設$X_1 \sim \chi^2(r_1)$，$X_2 \sim \chi^2(r_2)$且X_1，X_2獨立，則統計量

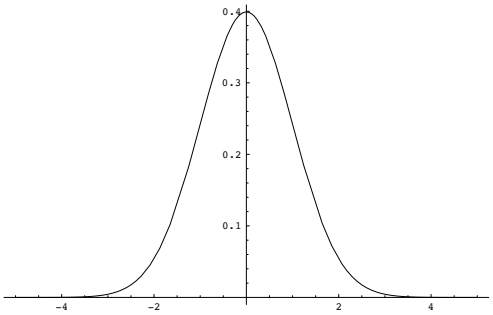
$$F = \frac{X_1/r_1}{X_2/r_2} \sim \mathcal{F}(r_1,\, r_2) \tag{1-14}$$

特性：

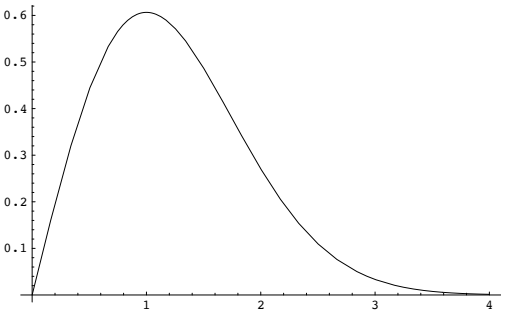
(1). 定義域爲$[0, \infty)$

(2). 不對稱，爲一右偏分配

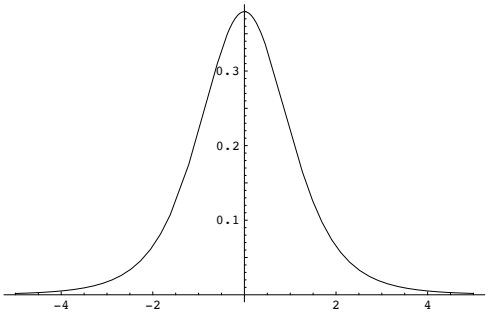
(3). 與卡方分配相似



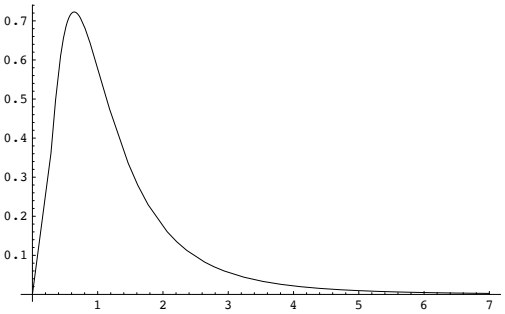
(a) 標準常態分配圖



(b) 卡方分配圖



(c) t分配圖



(d) $\mathcal{F}$分配圖

1.6 大數法則與中央極限定理

1 大數法則： 假設一隨機變數之機率密度函數$f(x)$，其平均數爲μ及變異數爲σ^2，且$0 < \sigma^2 < \infty$，令ε爲任意微小正數, 則

(1). 弱大數法則（Weak Law of Large Number）：

$$\lim_{n\to\infty} \mathrm{P}\left(\left|\overline{X} - \mu\right| > \varepsilon\right) = 0$$

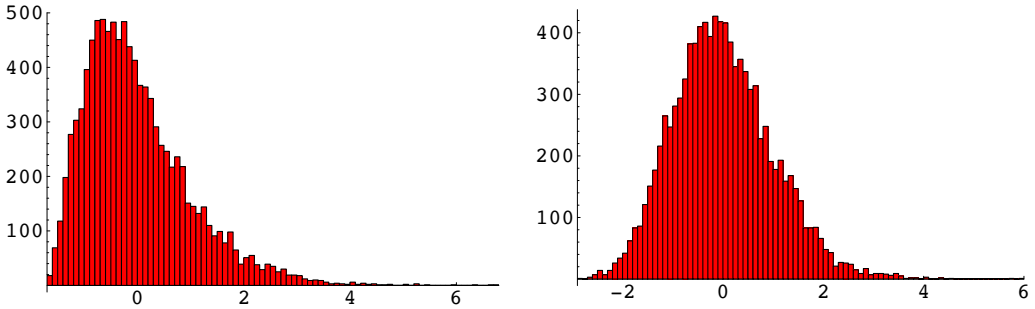
(2). 強大數法則（Strong Law of Large Number）：

$$\mathrm{P}\left(\left|\lim_{n\to\infty} \overline{X} - \mu\right| = 0\right) = 1$$

2 中央極限定理（Central Limit Theorem）： 若隨機變數$X_1, X_2, \ldots, X_n$爲抽自任意母體的一組隨機樣本。若 母體的期望值及變異數皆存在($\mu < \infty, \sigma^2 < \infty$), 則當樣本數趨近無窮大時($n \to \infty$)時,

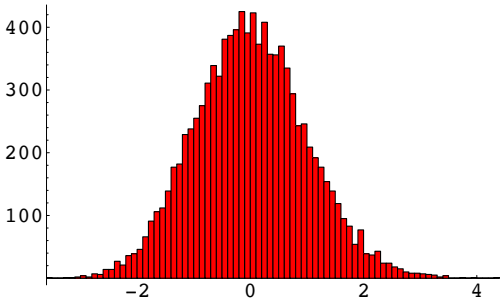
$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

一般來說, 當$n > 30$（即統計學上所謂之大樣本）時，上述定理即可成立。必須注意的是當μ或σ^2不存在時，此定理並不適用，例如柯西分配（Cauchy Distribution）。 下圖爲抽自Gamma(3,1)在不同樣本數下在$\overline{X}$的次數分配圖。由圖形可發現，實驗次數越多，經過標準化後之抽樣分配亦會愈趨近標準常態分配。

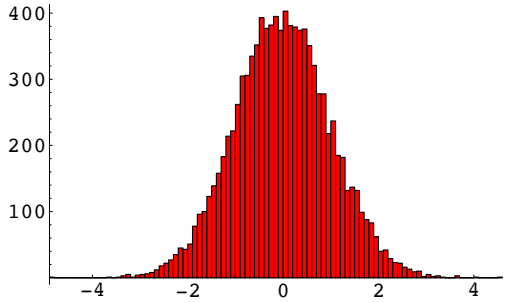


(a) $n = 1$

(b) $n = 10$



(c) $n = 100$



(d) $n = 1000$

1.7 常用的統計圖形

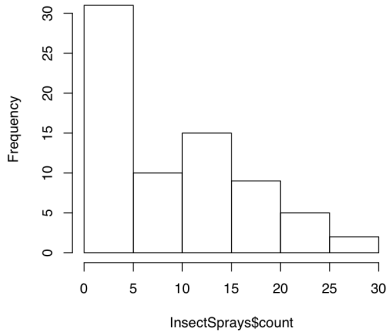
1 直方圖（Histogram plot）：可了解每個分數或組距發生的次數。通常適用於等距及比率變數。

2 盒形圖（Box plot）：其目的爲檢測盒子的對稱性；盒子的上緣爲Q_3，下緣爲Q_1，中間爲中位數。如果十分對稱，則可能是常態分佈。盒子外上下各有一線，上限爲$Q_3 + 1.5(Q_3 - Q_1)$，下限爲$Q_1 - 1.5(Q_3 - Q_1)$，檢測兩線之長度，表示偏度的大小。如果下線比上線長，則表示偏左分佈；如果上線比下線長，則表示偏右分佈。盒子愈大，則表示觀察體散佈愈大。由盒子上下界所延伸出的線，即稱之爲鬚（whiisker），是用以連接觀察體（非界外、極端值）的最大值與最小值。此外，盒鬚圖中亦可顯示出資料的界外值（ourlicrs，以“O”表示之）與極端值（extreme，以“E”表示之)。通常其界定是以任何觀察值與盒端的距離介於1.5至3倍盒長（即內四分位數$Q_3 - Q_1$）之間者，即稱之

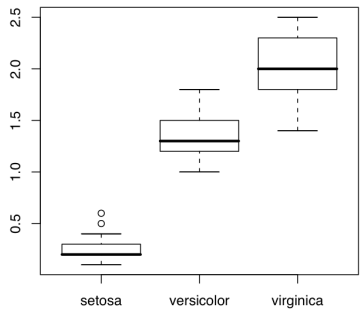
爲Outlier，而超過3倍的盒長者，即稱之爲Extreme。

3 散佈圖（Scatter plot）：用來說明某二個變量分佈的樣本的直角座標圖。如果數據大致落在一直線上，即表示兩變數具有高度線性相關；反之則否。

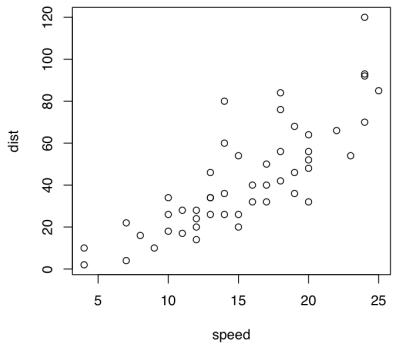
4 常態分位數圖（Q-Q plot）：以圖形來檢驗數據是否來自某一分配的方法。實際的應用裡，可利用常態分佈的Q-Q plot來做爲資料是否來自常態分佈的母體初步的檢驗。若樣本是來自常態分佈的母體，則其Q-Q圖近似於一條直線，否則便不是。



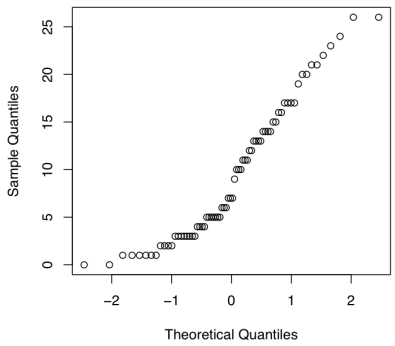
(a) 直方圖



(b) 盒形圖



(c) 散佈圖



(d) 常態分位數圖

教育程度學歷起薪盒型圖

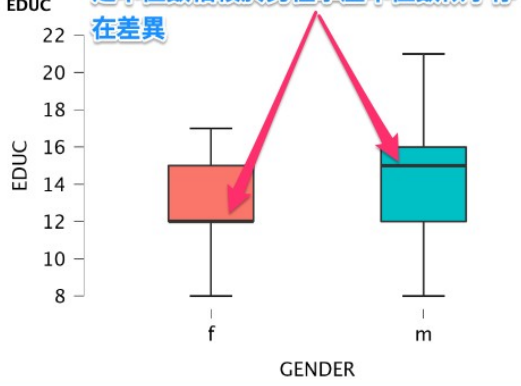
盒型圖，資料必需是屬量（等距、比率尺度）資料

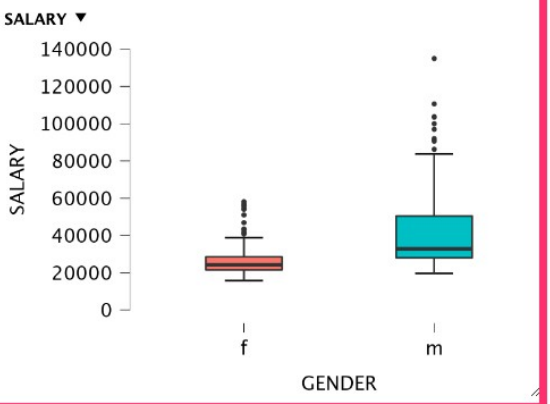
Descriptive Statistics

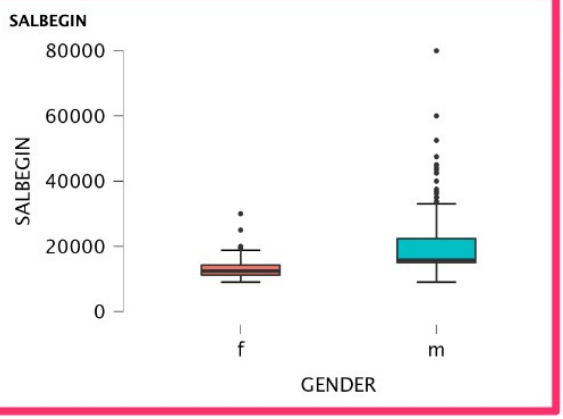
		Valid	Missing	Mean	Std. Deviation	Minimum	Maximum
EDUC	f	216	0	12.370	2.319	8.000	17.000
EDUC	m	258	0	14.430	2.979	8.000	21.000
SALARY	f	216	0	26031.921	7558.021	15750.000	58125.000
SALARY	m	258	0	41441.783	19499.214	19650.000	135000.000
SALBEGIN	f	216	0	13091.968	2935.599	9000.000	30000.000
SALBEGIN	m	258	0	20301.395	9111.781	9000.000	79980.000

Boxplots

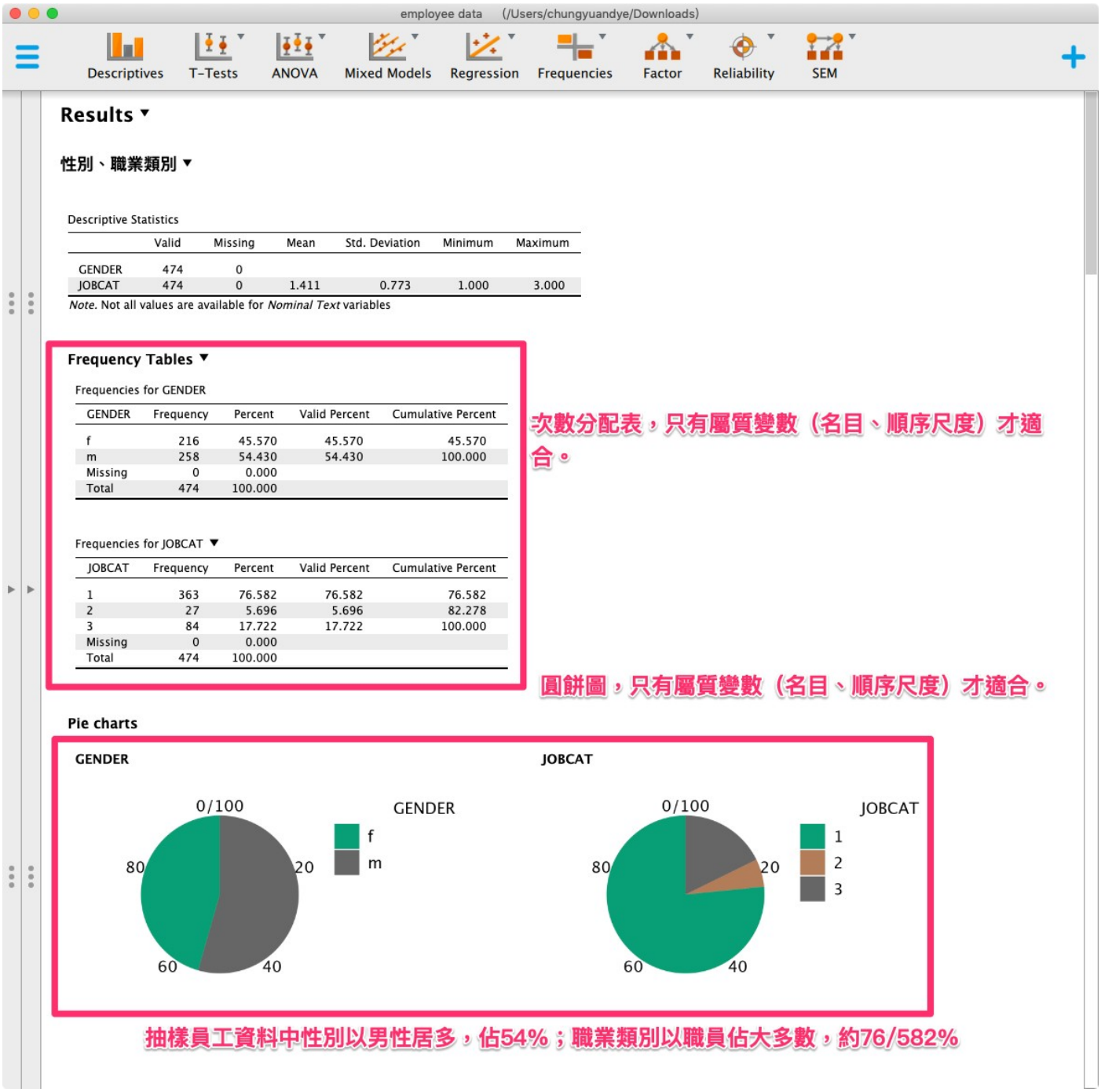
由盒型圖可發現男女性員工中女性學歷之中位數相較於男性學歷中位數似乎存在差異

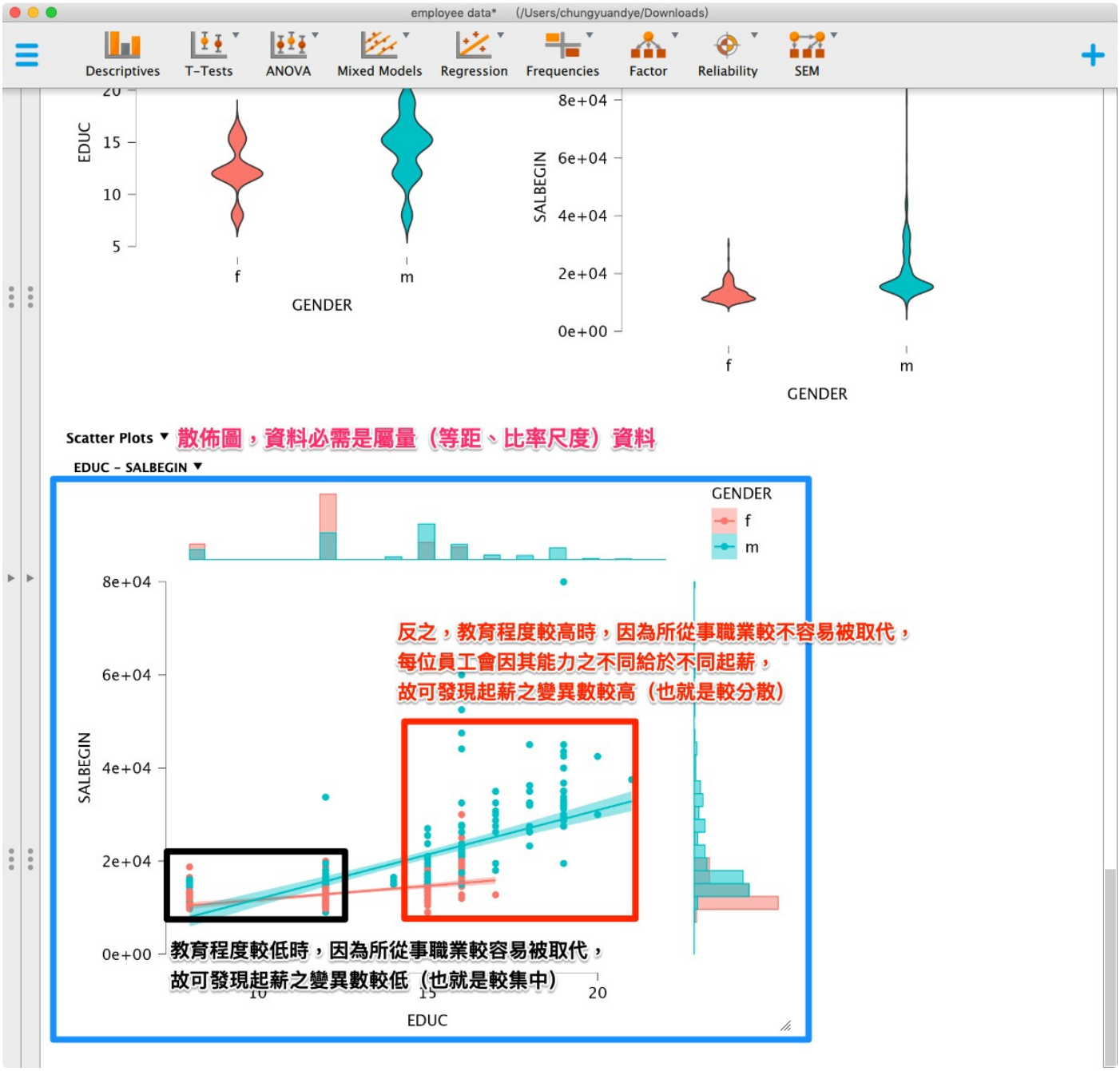






由盒型圖可發現男女性員工中，女性員工的起薪或目前薪資相較男性員工更為集中。


employee data (/Users/chungyuandye/Downloads)
Descriptives
T-Tests
ANOVA
Mixed Models
Regression
Frequencies
Factor
Reliability
SEM
Results
性別、職業類別
Descriptive Statistics
Valid Missing Mean Std. Deviation Minimum Maximum
GENDER 474 0
JOBCAT 474 0 1.411 0.773 1.000 3.000
Note. Not all values are available for Nominal Text variables
Frequency Tables
Frequencies for GENDER
GENDER Frequency Percent Valid Percent Cumulative Percent
f 216 45.570 45.570 45.570
m 258 54.430 54.430 100.000
Missing 0 0.000
Total 474 100.000
Frequencies for JOBCAT
JOBCAT Frequency Percent Valid Percent Cumulative Percent
1 363 76.582 76.582 76.582
2 27 5.696 5.696 82.278
3 84 17.722 17.722 100.000
Missing 0 0.000
Total 474 100.000
次數分配表，只有屬質變數（名目、順序尺度）才適合。
圓餅圖，只有屬質變數（名目、順序尺度）才適合。
Pie charts
GENDER
0/100
20
40
60
80
GENDER
f
m
JOBCAT
0/100
20
40
60
80
JOBCAT
1
2
3
抽樣員工資料中性別以男性居多，佔54%；職業類別以職員佔大多數，約76/582%

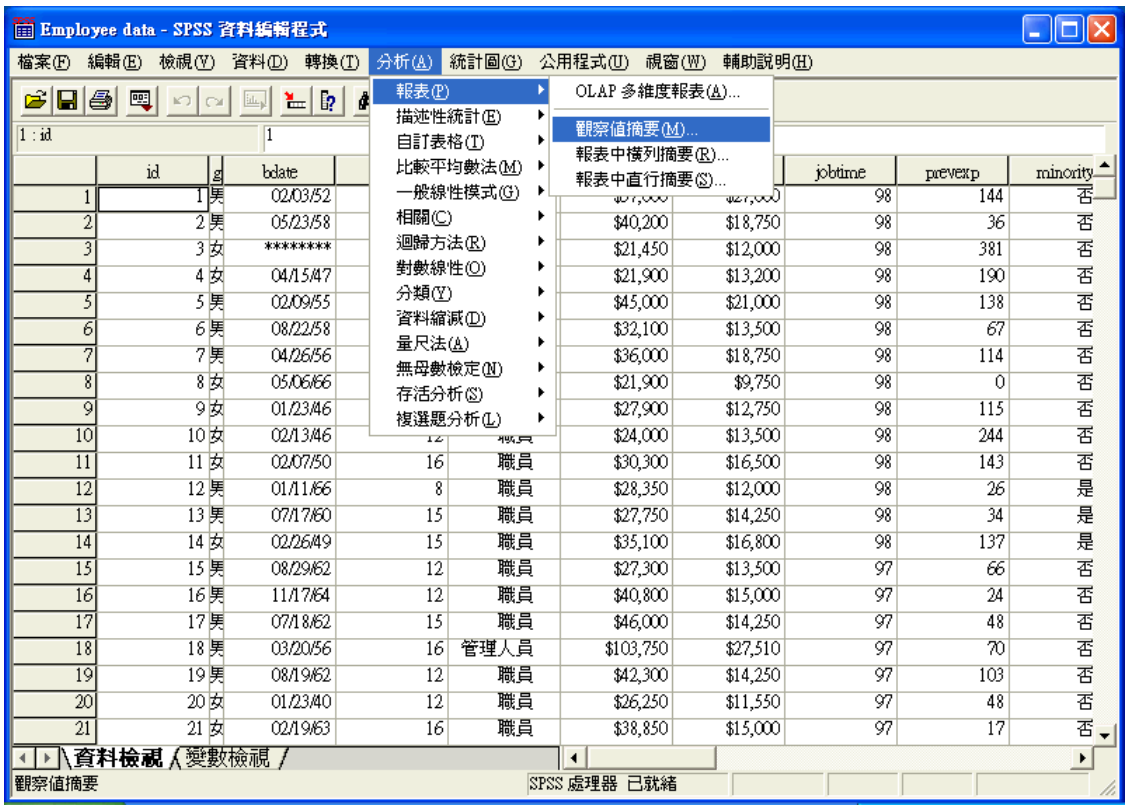

employee data* (/Users/chungyuandye/Downloads)
Descriptives
T-Tests
ANOVA
Mixed Models
Regression
Frequencies
Factor
Reliability
SEM
EDUC
GENDER
SALBEGIN
Scatter Plots
散佈圖，資料必需是屬量（等距、比率尺度）資料
EDUC - SALBEGIN
反之，教育程度較高時，因為所從事職業較不容易被取代，
每位員工會因其能力之不同給於不同起薪，
故可發現起薪之變異數較高（也就是較分散）
教育程度較低時，因為所從事職業較容易被取代，
故可發現起薪之變異數較低（也就是較集中）

1.8 SPSS操作範例

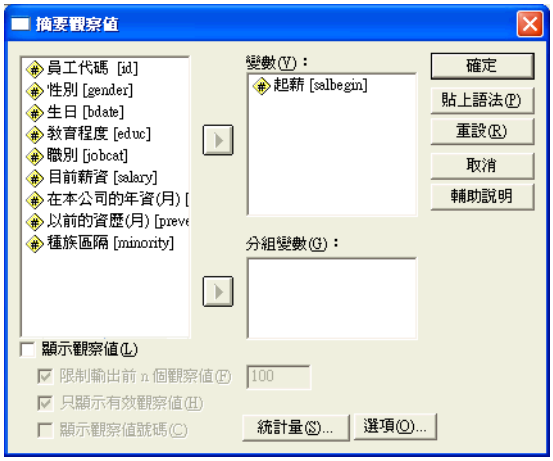
SPSS中資料的初步檢視方式有很多，例如：『觀察值摘要』、『次數分配表』、『描述統計量』及『預檢資料』，讀者可依據不同的需求分別進行。以Employee data.sav爲例，若想了解員工起薪資料或教育程度的分佈，依照上述分析方式介紹如下：

1. 觀察值摘要

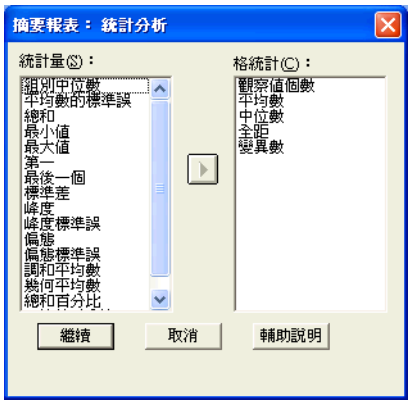
Step 1. SPSS功能列⇒分析⇒報表⇒觀察值摘要



Step 2. 將起薪移至變數方塊



Step 3. 選擇『統計量』，勾選欲輸出之統計量。



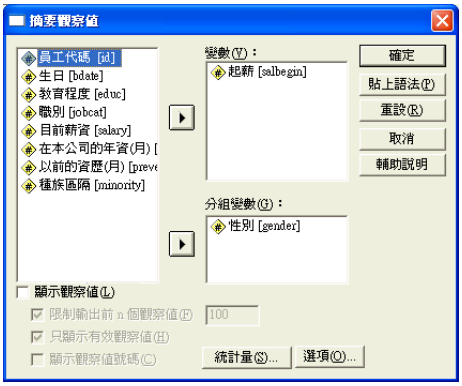
執行之後即可將報表輸出如下：

觀察值摘要

起薪

個數	平均數	中位數	範圍	變異數
474	17,016.09	15,000.00	70,980	61946944.959

當然也可以依照分組資料將報表輸出，例如想了解男女員工個別起薪資料的分佈，則只要將性別移至分組變數方塊中，即可輸出各組報表如下：



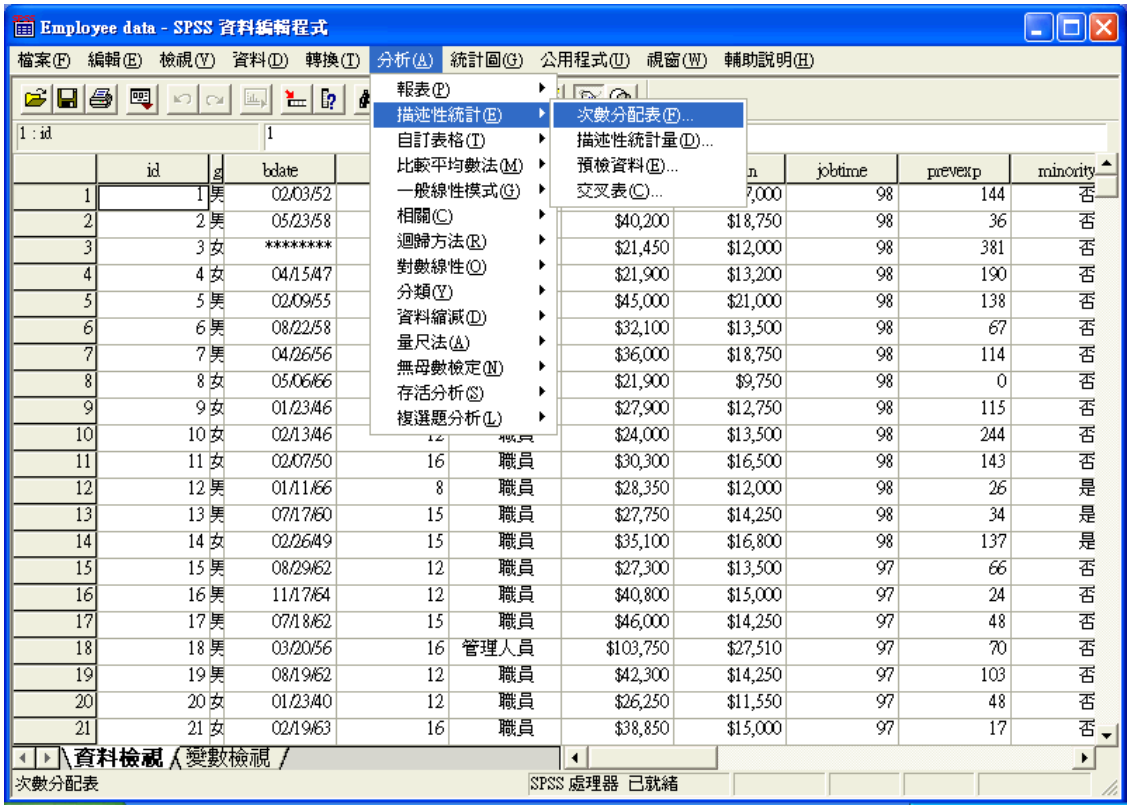
觀察值摘要

起薪

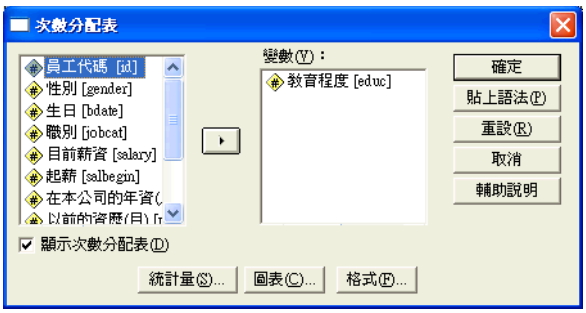
性別	個數	平均數	中位數	範圍	變異數
女	216	13,091.97	12,375.00	21,000	8617742.738
男	258	20,301.40	15,750.00	70,980	83024550.575
總和	474	17,016.09	15,000.00	70,980	61946944.959

2. 次數分配表

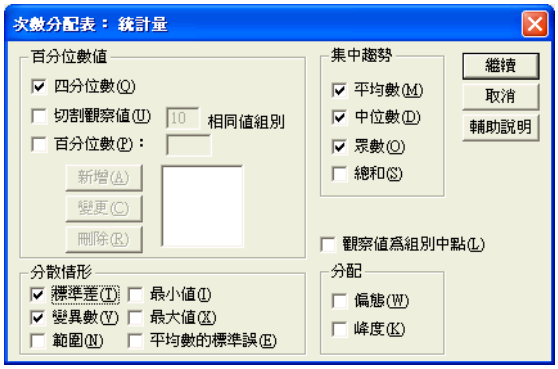
Step 1. SPSS功能列⇒分析⇒描述性統計⇒次數分配表



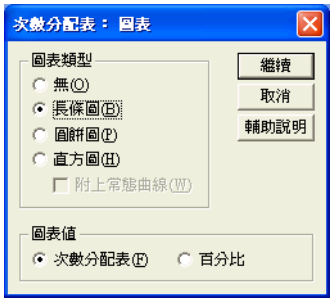
Step 2. 將教育程度移至變數方塊



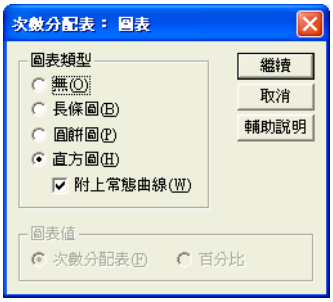
Step 3. 選擇『統計量』，勾選欲輸出之統計量。



Step 4. 選擇『圖表』，勾選欲輸出之圖表



(a)長條圖



(b)直方圖

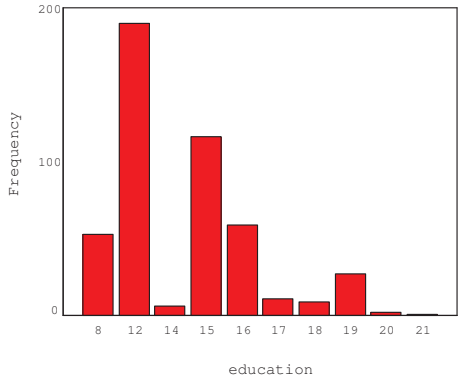
執行之後即可將報表輸出如下：

統計量

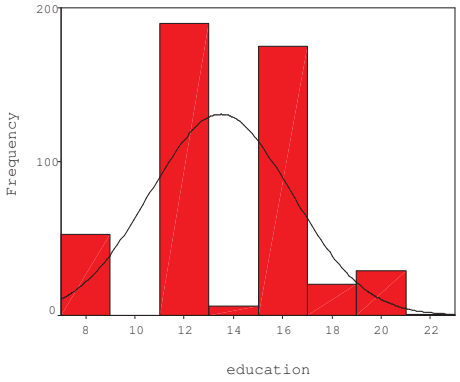
		教育程度
個數	有效的	474
	遺漏值	0
平均數		13.49
中位數		12.00
眾數		12
標準差		2.88
變異數		8.32
百分位數	25	12.00
	50	12.00
	75	15.00

教育程度

		次數	百分比	有效百分比	累積百分比
有效的	8	53	11.2	11.2	11.2
	12	190	40.1	40.1	51.3
	14	6	1.3	1.3	52.5
	15	116	24.5	24.5	77.0
	16	59	12.4	12.4	89.5
	17	11	2.3	2.3	91.8
	18	9	1.9	1.9	93.7
	19	27	5.7	5.7	99.4
	20	2	0.4	0.4	99.8
	21	1	0.2	0.2	100.0
	總和	474	100.0	100.0	



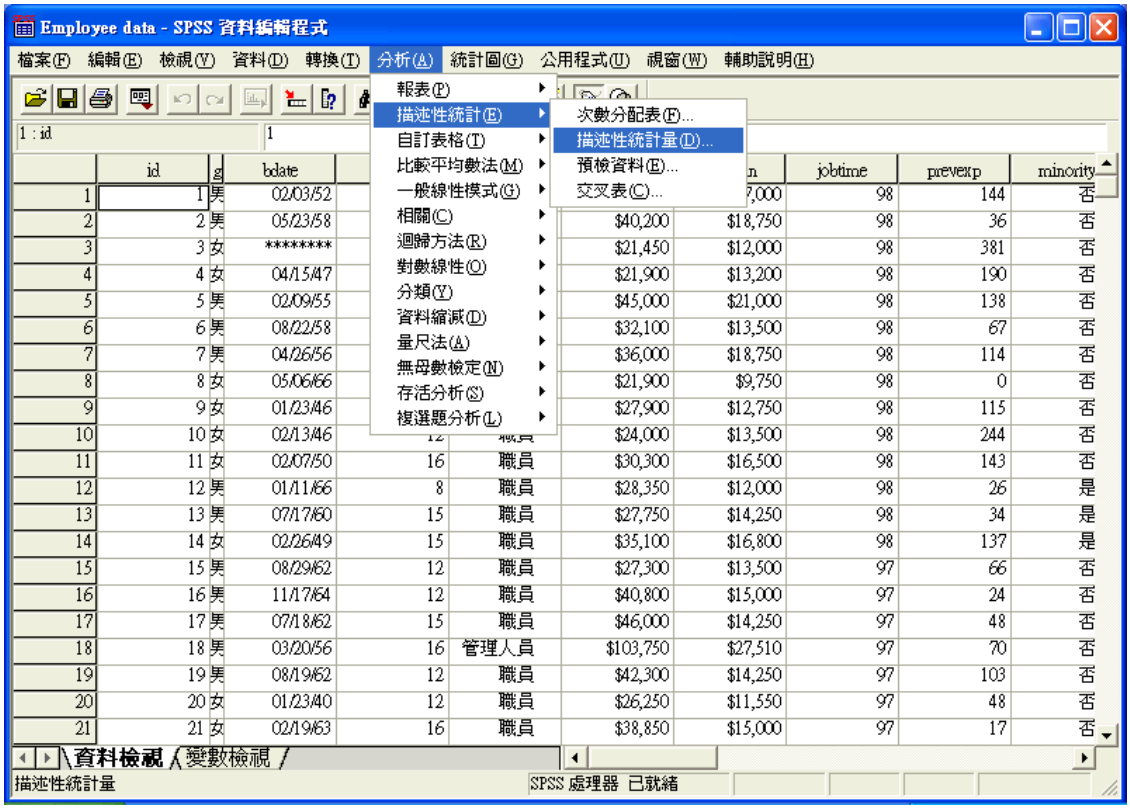
(a)教育程度條型圖



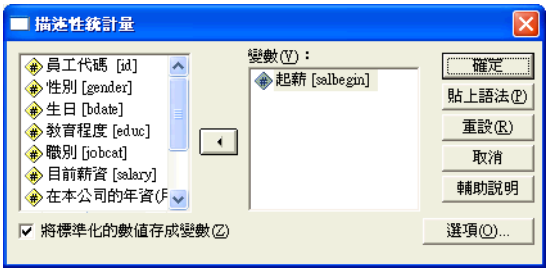
(b)教育程度直方圖

3. 描述性統計量

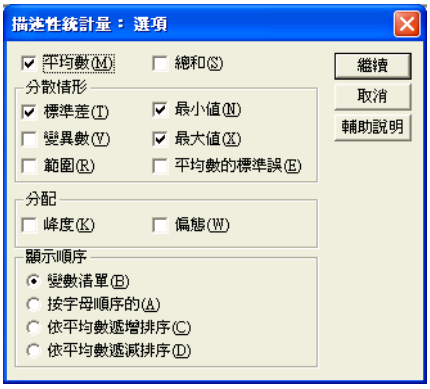
Step 1. SPSS功能列⇒分析⇒描述性統計⇒描述性統計量



Step 2. 將起薪移至變數方塊。若勾選『將標準化的數值存成變數』，則執行後會發現資料視窗將會多出一欄『起薪』之Z分數。



Step 3. 選擇『選項』，勾選欲輸出之統計量。



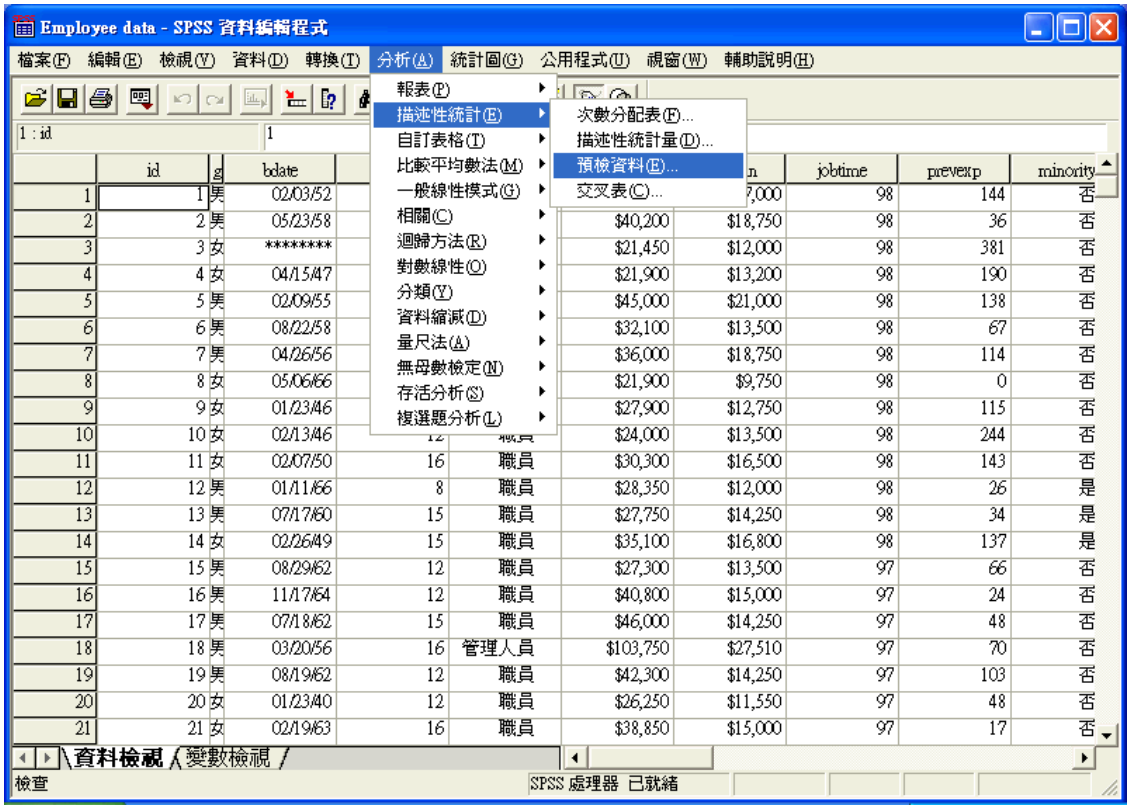
執行之後即可將報表輸出如下：

敘述統計

	個數	最小值	最大值	平均數	標準差
起薪	474	9,000	79,980	17,016.09	7,870.64
有效的 N (完全排除)	474				

4. **預檢資料**

Step 1. SPSS功能列⇒分析⇒描述性統計⇒預檢資料

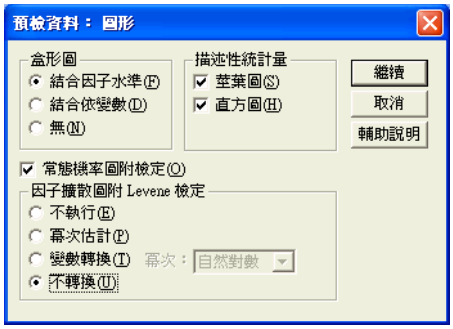


Step 2. 將起薪移至依變數方塊，並將性別移至因子[1]清單



[1] 調查中研究某一獨立變數或爲引起資料發生變動的原因。例如：價格對於銷售量之影響，價格即爲一因子。同樣地，研究四種不同電視節目對觀衆吸引力之影響，其中電視節目即爲一因子。

Step 3. 選擇『圖形』，勾選欲輸出之圖形及常態機率圖附Levene檢定[2]。



執行之後即可將報表輸出如下：

敘述統計

性別			統計量	標準誤
起薪 女	平均數		13,091.97	199.74
	平均數的 95% 信賴區間	下限	12,698.26	
		上限	13,485.67	
	刪除兩極端各 5% 觀察值之平均數		12,849.56	
	中位數		12,375.00	
	變異數		8617742.738	
	標準差		2,935.60	
	最小值		9,000	
	最大值		30,000	
	範圍		21,000	
	四分位全距		3,118.75	
	偏態		1.767	0.166
	峰度		5.352	0.330
男	平均數		20,301.40	567.27
	平均數的 95% 信賴區間	下限	19,184.30	
		上限	21,418.49	
	刪除兩極端各 5% 觀察值之平均數		19,249.94	
	中位數		15,750.00	
	變異數		83024550.575	
	標準差		9,111.78	
	最小值		9,000	
	最大值		79,980	
	範圍		70,980	
	四分位全距		7,687.50	
	偏態		2.390	0.152
	峰度		8.488	0.302

由以下Kolmogorov-Smirnov檢定[3]之報表可了解男女的起薪分佈在顯著水準0.05下並非服從常態分配。因此，在小樣本時，並不宜以SPSS中所提供之獨立樣本T檢定或單因子變異數分析檢定分析，應以兩獨立樣本之無母數統計方法檢定之。然而就本例而言，由於樣本數大，若想了解該公司男女之起薪是否相等，由中央極限定理可知，仍可利用SPSS中所提供之獨立樣本T檢定來檢定之。

[2] 關於Levene's test之檢定程序，讀者可至下列網址取得；
http://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm

[3] 關於Kolmogorov-Smirnov Goodness-of-Fit Test之檢定程序，讀者可至下列網址取得。
http://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm

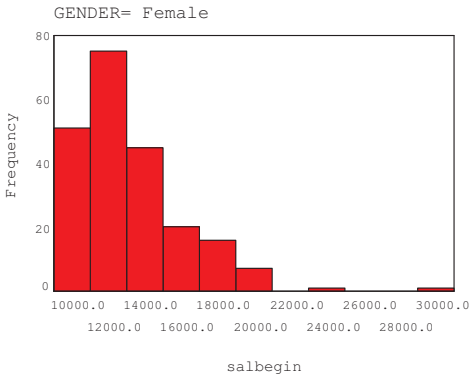
常態檢定

		Kolmogorov-Smirnov檢定		
	性別	統計量	自由度	顯著性
起薪	女	0.148	216	0.000
	男	0.259	258	0.000

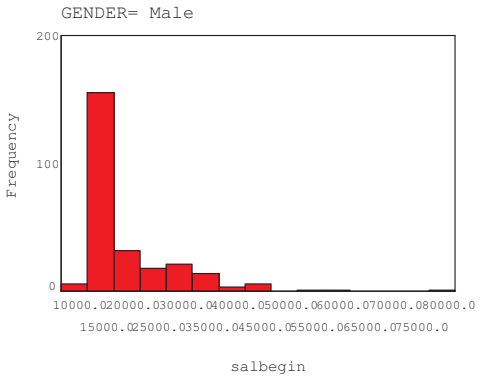
[a]Lilliefors 顯著性校正

變異數同質性檢定

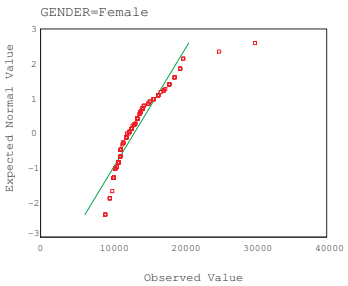
		Levene 統計量	分子自由度	分母自由度	顯著性
起薪	以平均數爲準。	105.969	1	472	0.000
	以中位數爲準。	33.917	1	472	0.000
	以中位數爲準，並含調整後的自由度	33.917	1	284.183	0.000
	以刪除極端值的平均數爲準	79.236	1	472	0.000



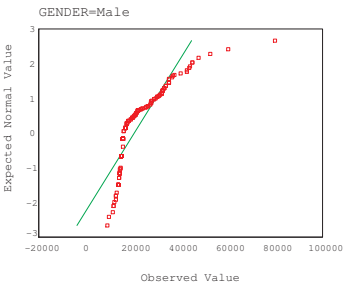
(a)女性



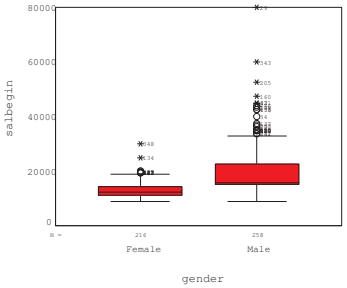
(b)男性



(a)女性



(b)男性



(c)盒形圖

第 2 章

區間估計與假設檢定

研究者欲了解母體的特徵，通常由研究的母體中抽取一部分樣本加以調查並蒐集資料。在樣本取得後，根據樣本所提供的資料對未知的母體參數做出適當的推論，這種方法我們稱爲推論統計學（Inferential Statistics）。而假設檢定與區間估計的建立則是推論統計學中重要的概念。

2.1 區間估計

區間估計乃指根據觀察值$x_1, x_2, \ldots, x_n$與運用抽樣分配及機率之原理，來預測一未知參數θ在某種程度下可能所在的範圍的方法。以機率形式表示爲

$$P(L \leq \theta \leq U) = 1 - \alpha \tag{2-1}$$

其中

1 $1 - \alpha$ 稱爲信賴水準（Confidence Level）

2 $(L,\ U)$ 稱爲θ的$(1 - \alpha) \times 100\%$信賴區間

3 L稱爲信賴區間的信賴下限

4 U稱爲信賴區間的信賴上限

由（2-1）可知，任何參數都有無限個信賴區間，即使信賴水準相同的情況下也不例外。信賴水準給定下，在應用上我們當然選擇在長度最小的信賴區間。

2.2 假設檢定

假設檢定乃指先給予母體未知參數一個假設值，在利用實驗之結果或母體抽出的樣本，應用機率理論去判斷此一假設值對或是不對，這種判斷的過程或這種方式的統計推論即稱爲假設檢定（Testing Hypotheses）。

2.2.1 名詞解釋

1 統計假設：任何有關於描述母體的敍述皆稱爲統計假設。

2 虛無假設（Null Hypotheses）：在假設檢定的術語中，希望被否定的統計假設稱爲虛無假設。

3 對立假設（Alternative Hypotheses）：在假設檢定的術語中，否定虛無假設而被認爲是對的統計假設稱爲對立假設。

4 簡單假設（Simple Hypothesis）：統計假設中只爲一個數值的假設，也就是說一個能完全地確定出機率分配的假設稱之。例如$H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$。

5 複合假設（Composite Hypothesis）：統計假設中，其假設不只包含一個數值，而是一個數值的集合。例如$H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$或$H_0 : \theta \geq \theta_0$ vs $H_1 : \theta < \theta_0$。

6 棄卻域（Reject Region）：樣本空間中導致否定虛無假設H_0之集合。

2.2.2 誤差之型態

利用樣本資料去估計母體參數時，由於抽樣可能產生偏差而可能產生錯誤決策的風險。

1 型I誤差 （Type I Error）：當虛無假設H_0爲眞，但檢定結果爲拒絕H_0稱之。

2 型II誤差（Type II Error）：當虛無假設H_0不眞，但檢定結果爲不拒絕H_0稱之。

3 α風險：型I誤差發生之機率；即$\alpha = P(\text{型I誤差}) = P(\text{拒絕}H_0 \mid H_0\text{爲眞})$。

4 β風險：型II誤差發生之機率；即$\beta = P(\text{型II誤差}) = P(\text{不拒絕}H_0 \mid H_0\text{爲僞})$。

5 顯著水準（Level of significance）：型I誤差發生最大的機率值，以α表示。

6 P-value：以樣本觀測值爲基準，做爲棄卻H_0爲錯誤決策的機率。

綜合以上討論，可得以下結論：

1 型I誤差與型II誤差互有關聯，其中之一的機率變大，則另一機率就變小。

2 若危險域的範圍愈小，則型I誤差α的機率就愈小，型II誤差β的機率就愈大；相反地，若危險域的範圍愈大，型I誤差的機率α就愈大，型II誤差β的機率就愈小。

3 增加樣本數n就可以使得α與β同時降低。

4 顯著水準是參數值爲虛無假設邊界點的α值。

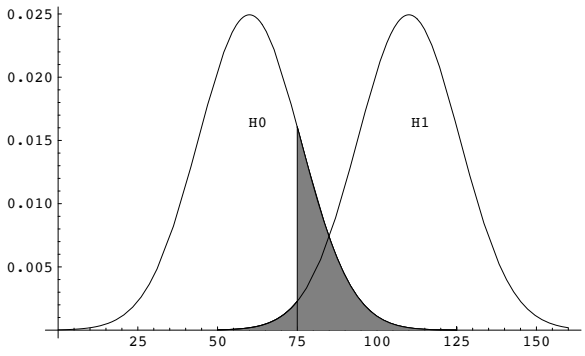


圖 2.1: α與β關係圖

2.3 假設檢定的型態與方法

假設檢定之型態有以下三種：

1 雙尾檢定：即雙尾對立假設的檢定問題，其統計假設爲

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

2 右尾檢定：即右尾對立假設的檢定問題，其統計假設爲

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta > \theta_0 \end{cases} \text{或} \begin{cases} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{cases}$$

3 左尾檢定：即左尾對立假設的檢定問題，其統計假設爲

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta < \theta_0 \end{cases} \text{或} \begin{cases} H_0 : \theta \geq \theta_0 \\ H_1 : \theta < \theta_0 \end{cases}$$

虛無假設H_0與對立假設H_1之關係如下：

1 均屬集合形式。

2 $\Theta_0 \cap \Theta_1 = \phi$，$\Theta_0 \cup \Theta_1 = \Theta$，其中$\Theta_0$表$\theta_0$之參數空間，$\Theta_1$表$\theta_1$之參數空間及$\Theta$表$\theta$之參數空間。

3 等號必存在Θ_0。

檢定的原理乃將虛無假設$H_0 : \theta = \theta_0$與經抽樣取得之樣本統計量$\widehat{\theta_0}$比較，若$|\theta_0 - \widehat{\theta_0}|$值愈小，即$H_0$爲正確的決策之機愈大。因此，在做相關決策時則必須有一比較之標準以決定棄卻H_0或不棄卻H_0，常用的方法有以下兩種：

1 利用檢定統計量與臨界值之比較，其決策法則爲當檢定統計量落入拒絕區，則棄卻H_0，否則不棄卻H_0。

2 利用P-value 與顯著水準 α 值之比較，決策法則爲當P-value $< \alpha$時，則棄卻H_0，否則不棄卻H_0[1]。

綜合以上說明可發現，除非具有足夠的證據可以棄卻H_0，否則我們只好接受。但在此處『接受』並不表示H_0爲眞，僅表示我們無充分證據可以來證明H_0是錯的。因此，我們並不以『接受H_0』做爲決策之結論。相對的，棄卻H_0時乃表示我們具有充分的證據可以拒絕，此時我們稱此一檢定稱爲具顯著性（Significance）。所以，統計假設檢定有時亦稱爲顯著性檢定（Significant Testing）。

[1] 在SPSS以及其他統計軟體檢定皆以此法爲主。爾後檢定僅需注意P-value是否大於α；當P-value $< \alpha$時，則棄卻H_0；否則不棄卻H_0。

2.4 母體平均數之檢定

在統計學的實務應用上，關於平均數的應用是較爲常見的，舉例來說品管工程師想了解一製程是否需要重新調整？抑或一教育學家想了解不同的教學方式對學生的學習能力是否有所差異？這些都必須經由平均數之比較才能得知結果。在本節我們首先介紹單一母體及兩母體之平均數比較之方法，分別介紹如下。

2.4.1 單一母體平均數之檢定

假設$X_1, X_2, \ldots, X_n$爲抽自$\mathcal{N}(\mu, \sigma^2)$的一組樣本數爲$n$的隨機樣本。若母體變異數$\sigma^2$已知，母體平均數之雙尾檢定方法介紹如下。首先我們欲檢定以下統計假設：

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

由於統計假設爲要檢定μ是否等於μ_0，因此若經抽樣所得的樣本平均數$\bar{x}$與μ_0差異甚大則傾向棄卻虛無假設H_0。由型I誤差之定義可知

$$\begin{aligned} \alpha &= P\left(\overline{X} < c_1 \text{ 或 } \overline{X} > c_2 \,\middle|\, \mu = \mu_0\right) \\ &= P\left(\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} < \frac{c_1 - \mu_0}{\sigma/\sqrt{n}}\right) + P\left(\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} > \frac{c_2 - \mu_0}{\sigma/\sqrt{n}}\right) \\ &= P\left(Z < \frac{c_1 - \mu_0}{\sigma/\sqrt{n}}\right) + P\left(Z > \frac{c_2 - \mu_0}{\sigma/\sqrt{n}}\right) \end{aligned}$$

由於爲雙尾檢定，故採左右尾之機率均爲$\alpha/2$。由圖2.2可知，斜線部分即爲棄卻域。 故可求得

$$-z_{\frac{\alpha}{2}} = \frac{c_1 - \mu_0}{\sigma/\sqrt{n}} \text{ 及 } z_{\frac{\alpha}{2}} = \frac{c_2 - \mu_0}{\sigma/\sqrt{n}}$$

或

$$c_1 = \mu_0 - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \text{ 及 } c_2 = \mu_0 + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

所以，棄卻域爲

$$C = \left\{ \left| \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \right| > z_{\frac{\alpha}{2}} \text{或} |\overline{x} - \mu_0| > z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right\} \tag{2-2}$$

故決策法則爲，當

$$\left| \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \right| > z_{\frac{\alpha}{2}} \text{或} |\overline{x} - \mu_0| > z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

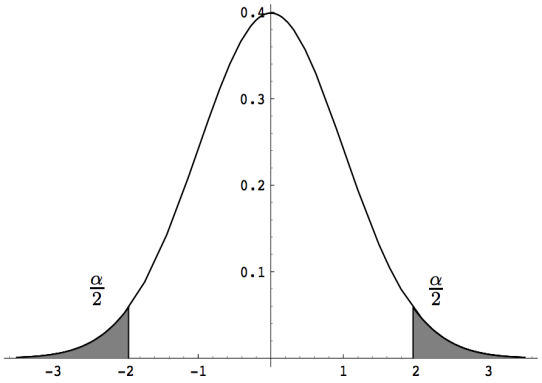


圖 2.2: 雙尾檢定圖

則棄卻虛無假設H_0：$\mu = \mu_0$，否則不棄卻H_0：$\mu = \mu_0$。

以上所介紹之檢定方法只在母體變異數σ^2已知的假設下成立。若母體變異數未知且樣本數$n < 30$，則需將樣本變異數S^2做爲σ^2之估計式。又

$$\frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t\,(n-1)$$

所以，棄卻域C可改寫爲

$$C = \left\{ \left| \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \right| > t_{\frac{\alpha}{2}}\,(n-1) \text{ 或 } |\overline{x} - \mu_0| > t_{\frac{\alpha}{2}}\,(n-1)\,\frac{s}{\sqrt{n}} \right\} \tag{2-3}$$

若以P-value檢定，則

$$\text{P-value} \quad = \quad P\left(|t| > \left| \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \right| \right)$$

必須注意的是在SPSS中並無母體變異數已知之單一母體平均數檢定，主要原因爲母體參數在實務上通常是未知；此外近年來電腦計算能力提昇，足以應付複雜的數值運算求得眞正的機率值。此外，在進行單一母體平均數檢定時，若要進行單尾平均數檢定，僅需將P-value除以2即可。

例 2-1. 爲了控制品質，品管工程師對於所生產罐裝調味料的重量是否都是16盎斯進行檢驗。因此，從生產線上隨機抽取了10個成品測量它們的重量如下：

重量：16.3　16.2　15.8　15.4　16.0　15.6　15.5　16.1　15.9　16.1

試在顯著水準$\alpha = 0.05$之下，試問罐裝調味料裝填機是否需要做調整？

解：我們欲檢定以下統計假設：

$$\begin{cases} H_0 : \mu = 16 \\ H_1 : \mu \neq 16 \end{cases}$$

由於$n = 10$，$\overline{x} = 15.890$，$s = 0.307$，所以棄卻域爲

$$C = \left\{ \left| \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \right| > t_{0.025}\,(9) \right\}$$

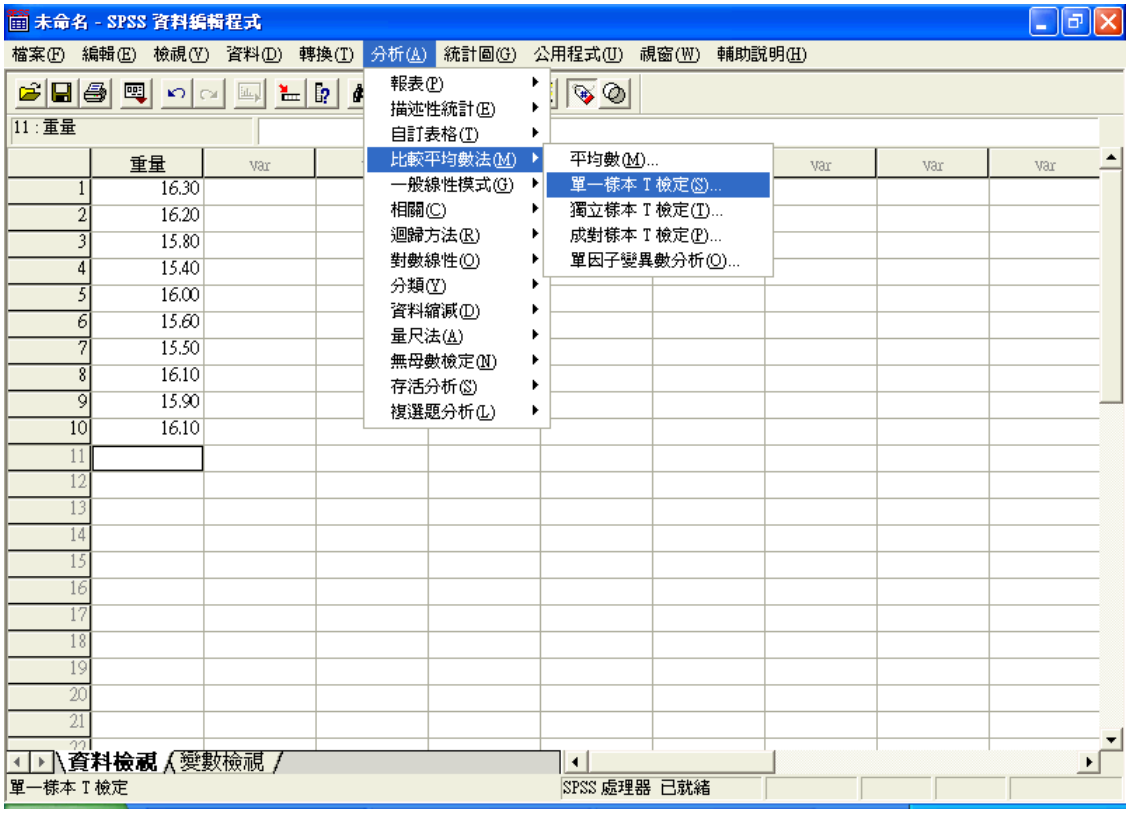
又

$$\left| \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \right| = \left| \frac{15.890 - 16}{0.307/\sqrt{10}} \right| = 1.133 < 2.262 = t_{0.025}\,(9)$$

所以不棄卻H_0，即無充分的證據顯示所生產的罐裝調味料的重量異於16盎斯。因此，機器不需調整。

SPSS之操作步驟如下：

Step 1. SPSS功能列⇒分析⇒比較平均數法⇒單一樣本T檢定



Step 2. 在檢定值文字方塊中輸入欲檢定值，此例檢定值爲16。



由SPSS輸出報表可知，由於P-value $= 0.287$，故不棄卻虛無假設

$$\begin{cases} H_0 : \mu = 16 \\ H_1 : \mu \neq 16 \end{cases}$$

因此，機器不需調整。

單一樣本檢定

	檢定值 = 16					
					差異的95%信賴區間	
	t	自由度	顯著性 (雙尾)	平均差異	下界	上界
重量	−1.133	9	0.287	−0.1100	−0.3297	0.1097

2.4.2 兩獨立母體平均數差之檢定

假設$X_1, X_2, \ldots, X_{n_1}$；$Y_1, Y_2, \ldots, Y_{n_2}$爲分別抽自$\mathcal{N}(\mu_x, \sigma_x^2)$及$\mathcal{N}(\mu_y, \sigma_y^2)$的兩組樣本數爲$n_1$及$n_2$的隨機樣本，$\mu_x - \mu_y$之檢定方法介紹如下：

首先，我們假設當σ_x^2，σ_y^2已知時，欲檢定以下問題

$$\begin{cases} H_0 : \mu_x - \mu_y = k \\ H_1 : \mu_x - \mu_y \neq k \end{cases}$$

由於統計假設爲要檢定$\mu_x - \mu_y$是否等於k，因此若經抽樣所得的樣本平均數$\bar{x} - \bar{y}$與k差異甚大則棄卻H_0虛無假設。由型I誤差之定義可知

$$\alpha = P\left(\overline{X} - \overline{Y} < c_1 \text{ 及 } \overline{X} - \overline{Y} > c_2 \,\middle|\, \mu_x - \mu_y = k\right)$$

$$= P\left(Z < \frac{c_1 - k}{\sqrt{\sigma_x^2/n_1 + \sigma_y^2/n_2}}\right) + P\left(Z > \frac{c_2 - k}{\sqrt{\sigma_x^2/n_1 + \sigma_y^2/n_2}}\right)$$

由於爲雙尾檢定，故採左右尾之機率均爲$\frac{\alpha}{2}$。故可求得

$$-z_{\frac{\alpha}{2}} = \frac{c_1 - k}{\sqrt{\sigma_x^2/n_1 + \sigma_y^2/n_2}} \text{ 及 } z_{\frac{\alpha}{2}} = \frac{c_2 - k}{\sqrt{\sigma_x^2/n_1 + \sigma_y^2/n_2}}$$

或

$$c_1 = k - z_{\frac{\alpha}{2}}\sqrt{\sigma_x^2/n_1 + \sigma_y^2/n_2} \text{ 及 } c_2 = k + z_{\frac{\alpha}{2}}\sqrt{\sigma_x^2/n_1 + \sigma_y^2/n_2}$$

所以，可求得棄卻域爲

$$C = \left\{ \left| \frac{\overline{x} - \overline{y} - k}{\sqrt{\sigma_x^2/n_1 + \sigma_y^2/n_2}} \right| > z_{\frac{\alpha}{2}} \text{或} |\overline{x} - \overline{y} - k| > z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2}} \right\} \tag{2-4}$$

當

$$\left| \frac{\overline{x} - \overline{y} - k}{\sqrt{\sigma_x^2/n_1 + \sigma_y^2/n_2}} \right| > z_{\frac{\alpha}{2}}$$

時棄卻虛無假設H_0：$\mu_x - \mu_y = k$，否則不棄卻虛無假設。然而，母體變異數在實務上通常未知。 所以母體變異數σ_x^2及σ_y^2未知且爲小樣本時，有下列兩種情況

1 若兩母體變異數相等($\sigma_x^2 = \sigma_y^2 = \sigma^2$)：

$$\frac{\overline{X} - \overline{Y} - k}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t_{\frac{\alpha}{2}}(n_1 + n_2 - 2) \tag{2-5}$$

其中S_p^2爲共同變異數σ^2之混和估計量（pooled estimator）

$$S_p^2 = \frac{(n_1 - 1)S_x^2 + (n_2 - 1)S_y^2}{n_1 + n_2 - 2} \tag{2-6}$$

故可求得棄卻域爲

$$C = \left\{ \left| \frac{\overline{x} - \overline{y} - k}{s_p\sqrt{1/n_1 + 1/n_2}} \right| > t_{\frac{\alpha}{2}}(n_1 + n_2 - 2) \right\} \tag{2-7}$$

所以，當

$$\left| \frac{\overline{x} - \overline{y} - k}{s_p\sqrt{1/n_1 + 1/n_2}} \right| > t_{\frac{\alpha}{2}}(n_1 + n_2 - 2)$$

時，棄卻虛無假設；否則不棄卻虛無假設。

2 若兩母體變異數不相等($\sigma_x^2 \neq \sigma_y^2$)：

$$\frac{\overline{X}-\overline{Y}-k}{\sqrt{S_x^2/n_1+S_y^2/n_2}} \sim t(v) \tag{2-8}$$

其中

$$v=\frac{\left(\frac{S_x^2}{n_1}+\frac{S_y^2}{n_2}\right)^2}{\frac{\left(\frac{S_x^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{S_y^2}{n_2}\right)^2}{n_2-1}} \tag{2-9}$$

故可求得棄卻域爲

$$C=\left\{\left|\frac{\overline{x}-\overline{y}-k}{\sqrt{S_x^2/n_1+S_y^2/n_2}}\right|>t_{\frac{\alpha}{2}}(v)\right\} \tag{2-10}$$

所以，當

$$\left|\frac{\overline{x}-\overline{y}-k}{\sqrt{S_x^2/n_1+S_y^2/n_2}}\right|>t_{\frac{\alpha}{2}}(v)$$

時，棄卻虛無假設；否則不棄卻虛無假設。

在SPSS中進行兩獨立母體平均數檢定時，程式會主動以Levene's test來檢定兩獨立母體之變異數是否相等，所以判讀報表時需先注意變異數檢定結果。Levene's test爲檢定多個母體變異數是否相等之統計方法，其統計假設及檢定統計量分別爲

$$\begin{cases} H_0:\sigma_1^2=\sigma_2^2=\cdots=\sigma_k^2 \\ H_1:\sigma_i^2\text{不全相等} \end{cases}$$

和

$$W=\frac{(N-k)\sum_{i=1}^{k} n_i\left(\overline{Z}_{i.}-\overline{Z}_{..}\right)^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i} n_i\left(Z_{ij}-\overline{Z}\right)^2} \sim F(k-1,N-k) \tag{2-11}$$

其中，N爲總樣本數，n_i爲各樣本之樣本數及$Z_{ij}=\left|X_{ij}-\overline{X}_i\right|$。當$W>F_\alpha(k-1,N-k)$時,棄卻虛無假設。

例 2-2. 將國民中學分爲都市國民中學與鄉村國民中學二類，各隨機抽取三年級學生10人作數學測驗，其成績如下：

都市國中（X）	73	59	69	66	55	87	89	89	84	75
鄉村國中（Y）	86	60	71	85	42	34	53	60	68	92

試檢定都市國中與鄉村國中三年級學生數學平均成績有顯著差異。（顯著水準=0.05，假定其成績之分配爲常態分配，而且二變異數相等）

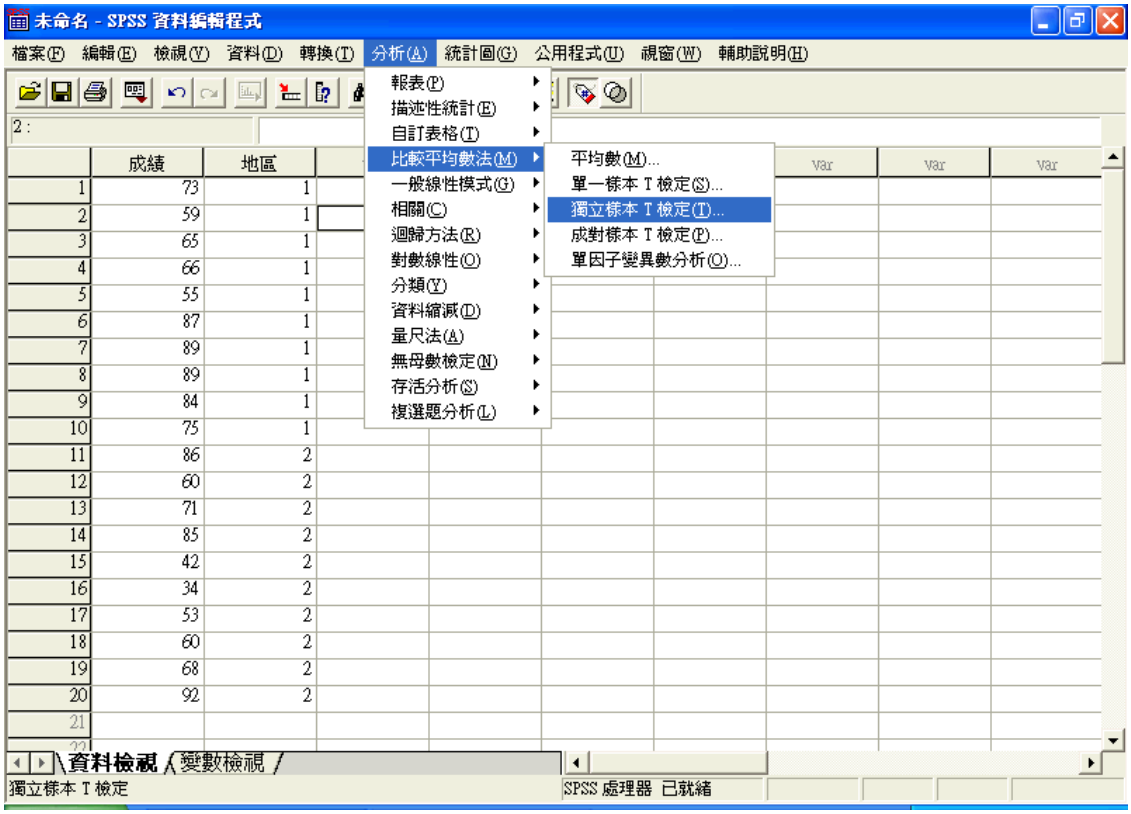
解： $\overline{x} = 74.6$, $\overline{y} = 65.1$, $s_x = 12.44$及$s_y = 19.15$，可求得棄卻域爲

$$C = \left\{ \left| \frac{\overline{x} - \overline{y}}{s_p \sqrt{1/n_1 + 1/n_2}} \right| > t_{\frac{\alpha}{2}}(18) = 2.101 \right\}$$

由於 $\left| \frac{\overline{x}-\overline{y}-0}{s_p\sqrt{1/n_1+1/n_2}} \right| = 1.316 < t_{0.025}(18) = 2.101$， 所以不棄卻虛無假設，即都市國中與鄉村國中三年級學生數學平均成績無顯著差異。

SPSS之操作步驟如下：

Step 1. SPSS功能列⇒分析⇒比較平均數法⇒獨立樣本T檢定



Step 2. 在組別選項中，分別輸入組別代號，本例組別代號都是國中爲『1』；鄉村國中爲『2』。



由SPSS輸出報表可知，由於在 Levene 檢定中，P-value $= 0.198 > 0.05$，故不棄卻虛無假設。又平均數檢定之P-value $= 0.205 > 0.05$，所以不棄卻虛無假設，即兩地區學生之成績並無顯著差異。

獨立樣本檢定

		Levene 檢定		t 檢定				
		F 檢定	顯著性	t值	自由度	顯著性(雙尾)	95% 信賴區間	
							下界	上界
成績	假設變異數相等	1.787	0.198	1.316	18	0.205	−5.67	24.67
	不假設變異數相等			1.316	15.449	0.207	−5.85	24.85

例 2-3. 以SPSS內部範例檔Employee data.sav爲例，試在$\alpha = 0.05$下，檢定某公司男女之起薪是否相等?

解：由SPSS所輸出之報表可知，由於兩母體之變異數檢定達顯著水準，故棄卻變異數相等之假設。又平均數檢定之P-value $= 0.000 < 0.05$，所以男女之起薪應有所差異。

獨立樣本檢定

		Levene 檢定		t 檢定				
		F 檢定	顯著性	t值	自由度	顯著性(雙尾)	95% 信賴區間	
							下界	上界
起薪	假設變異數相等	105.969	0.000	−11.152	472.000	0.000	−8479.7	−5939.16
	不假設變異數相等			−11.988	318.818	0.000	−8392.67	−6026.19

2.4.3 兩相依母體平均數差之檢定

若利用同樣的受試者參加不同實驗所得之資料（例如不同的受測者在某項藥物治療前與治療後之體重），由於樣本抽取的機率爲非獨立而是相依的情況，故稱爲相依母體。假設$X_1, X_2, \ldots, X_{n_1}$；$Y_1, Y_2, \ldots, Y_{n_2}$爲分別爲實驗前與實驗後之樣本資料，若資料爲來自常態母體。令$\mu_d = \mu_x - \mu_y$，則當σ_d已知情況下，μ_d之檢定方法介紹如下。我們欲檢定以下統計假設：

$$\begin{cases} H_0 : \mu_d = k \\ H_1 : \mu_d \neq k \end{cases}$$

由於統計假設爲要檢定μ_d是否等於k，因此若經抽樣所得的樣本平均數$\overline{d}$（$d_i = x_i - y_i$, $\overline{d} = \sum_{i=1}^{n} d_i/n$）與$k$差異甚大則棄卻$H_0$虛無假設。由型I誤差之定義可知

$$\begin{aligned}\alpha &= P\left(\overline{d} < c_1 \text{ 或 } \overline{d} > c_2 \,\middle|\, \mu_d = k\right) \\ &= P\left(Z < \frac{c_1 - k}{\sigma_d/\sqrt{n}}\right) + P\left(Z > \frac{c_2 - k}{\sigma_d/\sqrt{n}}\right)\end{aligned}$$

由於爲雙尾檢定，故採左右尾之機率均爲$\alpha/2$。故可求得

$$c_1 = k - z_{\frac{\alpha}{2}}\frac{\sigma_d}{\sqrt{n}} \text{ 及 } c_2 = k + z_{\frac{\alpha}{2}}\frac{\sigma_d}{\sqrt{n}}$$

所以，棄卻域爲

$$C = \left\{\left|\frac{\overline{d} - k}{\sigma_d/\sqrt{n}}\right| > z_{\frac{\alpha}{2}}\right\} \tag{2-12}$$

即當

$$\left|\frac{\overline{d} - k}{\sigma_d/\sqrt{n}}\right| > z_{\frac{\alpha}{2}} \text{或} \left|\overline{d} - k\right| > z_{\frac{\alpha}{2}}\frac{\sigma_d}{\sqrt{n}}$$

則棄卻虛無假設H_0：$\mu_d = k$，否則不棄卻H_0：$\mu_d = k$。

以上所介紹之檢定方法只在母體變異數σ_d^2已知的假設下成立。若母體變異數未知且樣本數$n < 30$，則需將樣本變異數S_d^2做爲σ_d^2之估計式。又

$$\frac{\overline{d} - k}{S_d/\sqrt{n}} \sim t\,(n-1)$$

所以，棄卻域C可改寫爲

$$C = \left\{\left|\frac{\overline{d} - k}{s_d/\sqrt{n}}\right| > t_{\frac{\alpha}{2}}\,(n-1)\right\} \tag{2-13}$$

即當

$$\left|\frac{\overline{d} - k}{s_d/\sqrt{n}}\right| > t_{\frac{\alpha}{2}}(n-1)$$

則棄卻虛無假設H_0：$\mu_d = k$，否則不棄卻H_0：$\mu_d = k$。

例 2-4. 某研究學者欲瞭解課後補習能否增進學習能力，以隨機方式抽取12位學生，補習前後之成績如下：

補習前	52	61	58	57	59	62	56	57	61	58	55	60
補習後	67	59	71	58	70	67	66	60	66	61	66	68

試以$\alpha = 0.05$，檢定補習前後成績是否不同？

解：我們欲檢定以下統計假設：

$$\begin{cases} H_0 : \mu_d = 0 \\ H_1 : \mu_d \neq 0 \end{cases}$$

由於$\overline{d} = -6.92$及$s_d = 4.99$，可求得棄卻域即檢定統計量分別爲

$$C = \left\{ \left| \frac{\overline{d} - 0}{s_d \sqrt{12}} \right| > t_{0.025}\,(11) = 2.201 \right\}$$

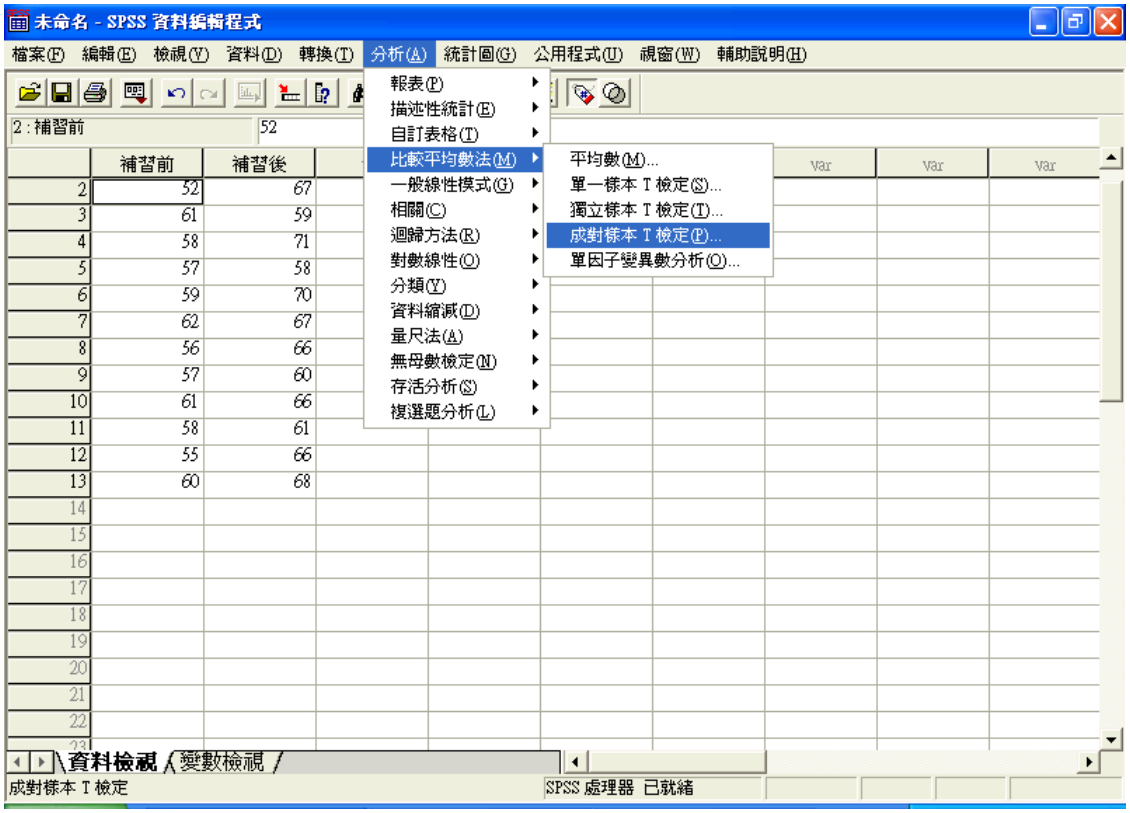
又

$$\left| \frac{\overline{d}}{s_d / \sqrt{12}} \right| = 5.001 > t_{0.025}(11) = 2.201$$

所以棄卻虛無假設$H_0 : \mu_d = 0$，即補習先後成績的確有所差異。

SPSS之操作步驟如下：

Step 1. SPSS功能列⇒分析⇒比較平均數法⇒配對樣本T檢定



Step 2. 選擇配對變數



由SPSS輸出報表可知，由於P-value $= 0.000 < 0.05$，所以棄卻虛無假設，即補習前後成績的卻有所差異。

成對樣本檢定

	成對變數差異				t	自由度	顯著性 (雙尾)
	平均數	標準差	差異的95%信賴區間				
			下界	上界			
補習前 - 補習後	−6.92	4.99	−9.94	−3.91	−5.001	12	0.000

第 3 章

卡方檢定

當我們檢定母體平均數時均假設資料爲來自常態母體。然而母體是否眞是如此？在統計學裡，我們必需要有一套嚴謹的方法來檢驗，以驗證理論分配是否正確。本章要處理的統計問題都具有一個共同特色，都是將統計資料分類或分組，計算每一類（組）發生的次數，再利用卡方分配來進行假設檢定，故此類檢定方法因而稱爲卡方檢定。針對不同類別的統計資料及有興趣探討的問題，卡方檢定應用有適合度檢定、獨立性檢定及齊一性檢定三種。

3.1 適合度檢定

適合度檢定爲檢定一組隨機樣本是否來自某一特定分配，統計假設爲

$$\begin{cases} H_0：\text{資料服從某特定分配} \\ H_1：\text{資料不服從某特定分配} \end{cases}$$

其原理爲藉由比較樣本觀察值的次數分配和假設的母體分配是否「接近」或「相似」，來判斷資料是否來自假設的分配。 此分類資料常以次數分配表的方式呈現資料。先將N個觀察值的樣本分成k組後，計算各組的觀察值個數O_i。若各組之觀察值O_i與其相對應之理論次數E_i接近，則可推論資料爲來自假設的分配。其檢定統計量爲

$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(k-1-p) \tag{3-1}$$

其中p爲未知母體參數個數。由上式可知，當H_0不爲眞時，檢定統計量會變大。因此，當檢定統計量落入棄卻域

$$C = \left\{ \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} > \chi^2_\alpha(k-1-p) \right\}$$

時，則應棄卻虛無假設H_0。

例 3-1. 假設5年前台北市的上班族上班的交通工具如下：

上班方式	坐公車	自己開車	騎機車	坐交通車	坐火車	其他
所佔比例	58%	12%	15%	6%	3%	6%

現在市政府委託一民間機構調查台北市民上班的方式是否有所改變，經該機構調查800位台北市的上班族，得資料如下：

上班方式	坐公車	自己開車	騎機車	坐交通車	坐火車	其他
人數	364	186	198	28	14	10

試在$\alpha = 0.05$下，檢定台北市民上班的方式是否有所改變？

解：我們所欲檢定問題如下：

$$\begin{cases} H_0：台北市民上班的方式沒改變 \\ H_1：台北市民上班的方式有改變 \end{cases}$$

資料經整理得

上班方式	坐公車	自己開車	騎機車	坐交通車	坐火車	其他
觀察人數	364	186	198	28	14	10
理論次數	464	96	120	48	24	48
$\frac{(O_i-E_i)^2}{E_i}$	13.79	84.38	50.70	8.33	4.17	30.08

因爲

$$\sum_{i}^{6} \frac{(O_i - E_i)^2}{E_i} = 191.45 > \chi^2_{0.05}(5) = 11.07$$

故棄卻虛無假設，即上班的方式有顯著改變。

例 3-2. 529株碗豆，具AA、Aa、aa各有138、265、126株，是否違反1：2：1比例?

解： 我們所欲檢定問題如下：

$$\begin{cases} H_0：AA\colon Aa\colon aa = 1\colon 2\colon 1 \\ H_1：AA\colon Aa\colon aa \neq 1\colon 2\colon 1 \end{cases}$$

資料經整理得

豌豆基因	AA	Aa	aa
觀察次數	138	265	126
理論次數	132.25	264.5	132.25
$\frac{(O_i-E_i)^2}{E_i}$	0.25	0.0009	0.30

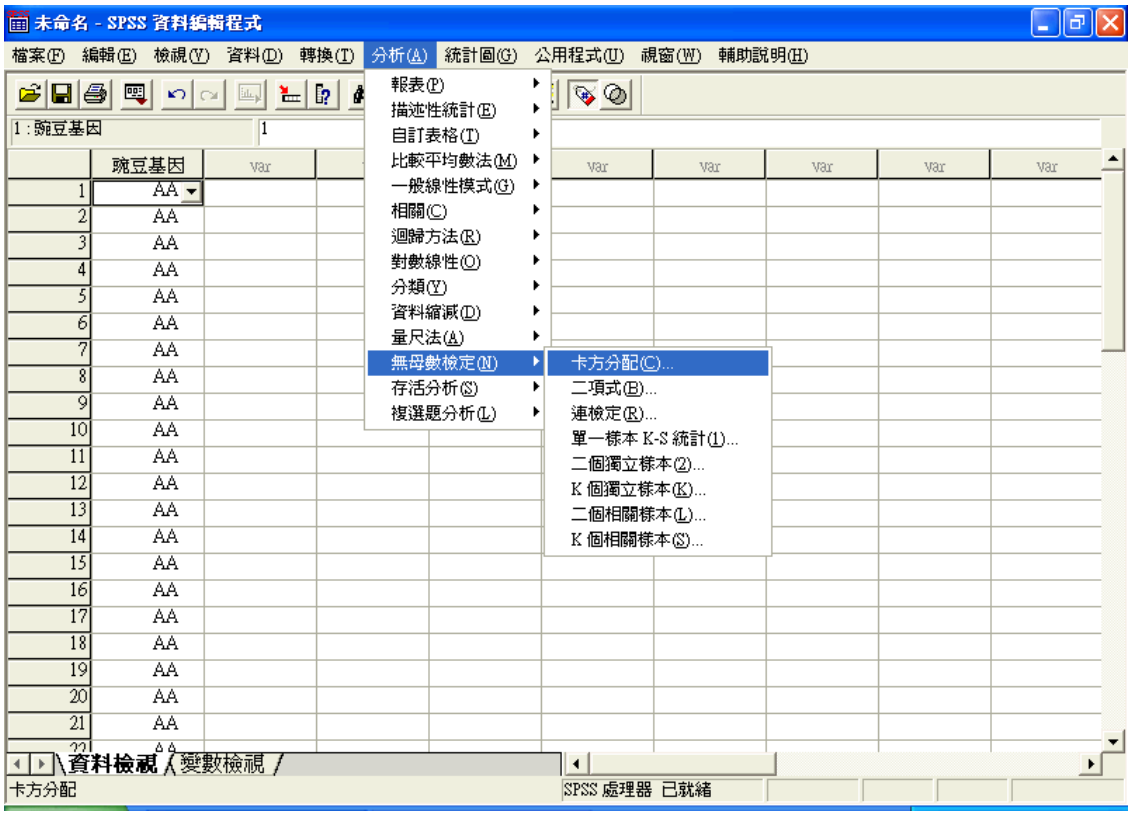
因爲

$$\sum_{i=1}^{3}\frac{(O_i-E_i)^2}{E_i}=0.5509<\chi^2_{0.05}(2)=5.99$$

所以不棄卻虛無假設，即豌豆基因比例爲$AA:\ Aa:\ aa=1:2:1$。

SPSS之操作步驟如下：

Step 1. SPSS功能列⇒分析⇒無母數檢定⇒卡方分配



Step 2.



由SPSS輸出報表可知，因爲P-value $=$ 0.761 $>$ 0.05，所以不棄卻虛無假設，即豌豆基因比例爲AA：Aa：aa=1：2：1。

豌豆基因

	觀察個數	期望個數	殘差
AA	138	132.3	5.8
Aa	265	264.5	0.5
aa	126	132.3	−6.3
總和	529		

檢定統計量

	豌豆基因
卡方	0.546
自由度	2
漸近顯著性	0.761

需注意的是，卡方適合度檢定通常只適用於大樣本。一般說來，當所有的$E_i \geq 5$時，卡方適合度檢定才適用。若有$E_i < 5$時，該組必須與相鄰組合併至$E_i \geq 5$爲止，即組數亦必須減少。

3.2 獨立性檢定

在資料分析中的兩個變數代表兩種不同的分類方法，資料經過兩種不同的分類方法所觀察到的結果，通常可交叉整理成列聯表 。獨立性檢定爲檢定兩個名目變數之間是否有相互影響。例如：性別（男或女）與抽煙行爲（抽或不抽）是否有關？性別（男或女）與素食行爲（葷食、素食）是否有關？學歷（大學、高中、國

表 3.1: $r \times c$列聯表

		B				
		B_1	B_2	$\cdots$	B_c	總計
A	A_1	O_{11}	O_{12}	$\cdots$	O_{1c}	$O_{1.}$
	A_2	O_{21}	O_{22}	$\cdots$	O_{2c}	$O_{2.}$
		$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
	A_r	O_{r1}	O_{r2}	$\cdots$	O_{rc}	$O_{r.}$
	總計	$O_{.1}$	$O_{.2}$	$\cdots$	$O_{.c}$	$O = n$

中、國小）與宗教信仰（佛教、基督教、天主教、道教）是否有關？統計假設爲

$$\begin{cases} H_0：\text{列變數與行變數獨立} \\ H_1：\text{列變數與行變數不獨立} \end{cases}$$

若虛無假設正確，則樣本資料在列聯表中之期望次數爲行總和與列總和的乘積再除以樣本大小，即

$$E_{ij} = n \times \frac{O_{i.}}{n} \times \frac{O_{.j}}{n}, i = 1, 2, \ldots r; j = 1, 2, \ldots, c \tag{3-2}$$

因此，若各組之觀察值O_{ij}與其相對應之理論次數E_{ij}接近，則可推論列變數與行變數獨立。其檢定統計量爲

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}} \sim \chi^2\left((r-1)(c-1)\right) \tag{3-3}$$

由上式可知，當H_0不爲眞時，檢定統計量會變大，所以當檢定統計量落入棄卻域

$$C = \left\{ \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}} > \chi^2_{\alpha}\left((r-1)(c-1)\right) \right\} \tag{3-4}$$

時，則應棄卻虛無假設H_0。SPSS在此也會提供概似比檢定統計量

$$2\sum_{i=1}^{r}\sum_{j=1}^{c}O_{ij} \times \log\frac{Oij}{E_{ij}} \sim \chi^2\left((r-1)(c-1)\right) \tag{3-5}$$

所以當檢定統計量落入棄卻域

$$C = \left\{ 2\sum_{i=1}^{r}\sum_{j=1}^{c}O_{ij} \times \log\frac{Oij}{E_{ij}} > \chi^2_{\alpha}\left((r-1)(c-1)\right) \right\} \tag{3-6}$$

時，則應棄卻虛無假設H_0。

與卡方適合度檢定相同，卡方獨立性檢定通常只適用於大樣本，一般說來當所有的$E_{ij} \geq 5$時，卡方適合度檢定才適用。其主要原因爲若有$E_{ij} < 5$時，則容易產生異常大的卡方檢定統計量，因此該行或該列分類則建議與相鄰行或列合併至$E_{ij} \geq 5$爲止。然而，合併資料組別雖然可以方便利用卡方檢定，卻容易讓資料的特徵消失以及因變項減少而失去解釋意義。實務上在處理大型的列聯表時，只要期望值小於5的細格數比例不太高且而所有的$E_{ij} \geq 1$，即使有細格（O_{ij}）爲0（原則上細格爲0的比例不超過20%），卡方檢定亦適用。 必須注意的是，當自由度爲1時，(3.3)式必須作Yates'連續性校正（Yates' correction for continuity）以避免誤差過大，即

$$\sum_{i=2}^{r}\sum_{j=1}^{2}\frac{(|O_{ij}-E_{ij}|-0.5)^2}{E_{ij}} \sim \chi^2(1) \tag{3-7}$$

另一方面，當樣本數不大或有細格爲0的比例超過所有20%時，則建議採用Fisher's exact test作爲兩個類別變項之相關的檢定方法，其統計量爲

$$p = \sum_{x=O_{11}}^{O_{\cdot 1}} \frac{\binom{O_{1\cdot}}{x}\binom{O_{2\cdot}}{O_{\cdot 1}-x}}{\binom{n}{O_{\cdot 1}}} \tag{3-8}$$

若$p < \alpha$時，則棄卻虛無假設。

例 3-3. 以SPSS內部範例檔voter.sav爲例，試在$\alpha = 0.05$下，檢定1992年美國總統大選選民性別與投票傾向是否有關？

解：我們所欲檢定問題如下：

$$\begin{cases} H_0：\text{選民性別與投票傾向無關} \\ H_1：\text{選民性別與投票傾向有關} \end{cases}$$

SEX * VOTE

			VOTE			總和
			Bush	Perot	Clinton	
SEX	male	個數	315	152	337	804
		期望個數	287.7	121.0	395.3	804.0
		總和的 %	17.1%	8.2%	18.2%	43.5%
	female	個數	346	126	571	1043
		期望個數	373.3	157.0	512.7	1043.0
		總和的 %	18.7%	6.8%	30.9%	56.5%
總和		個數	661	278	908	1847
		期望個數	661.0	278.0	908.0	1847.0
		總和的 %	35.8%	15.1%	49.2%	100.0%

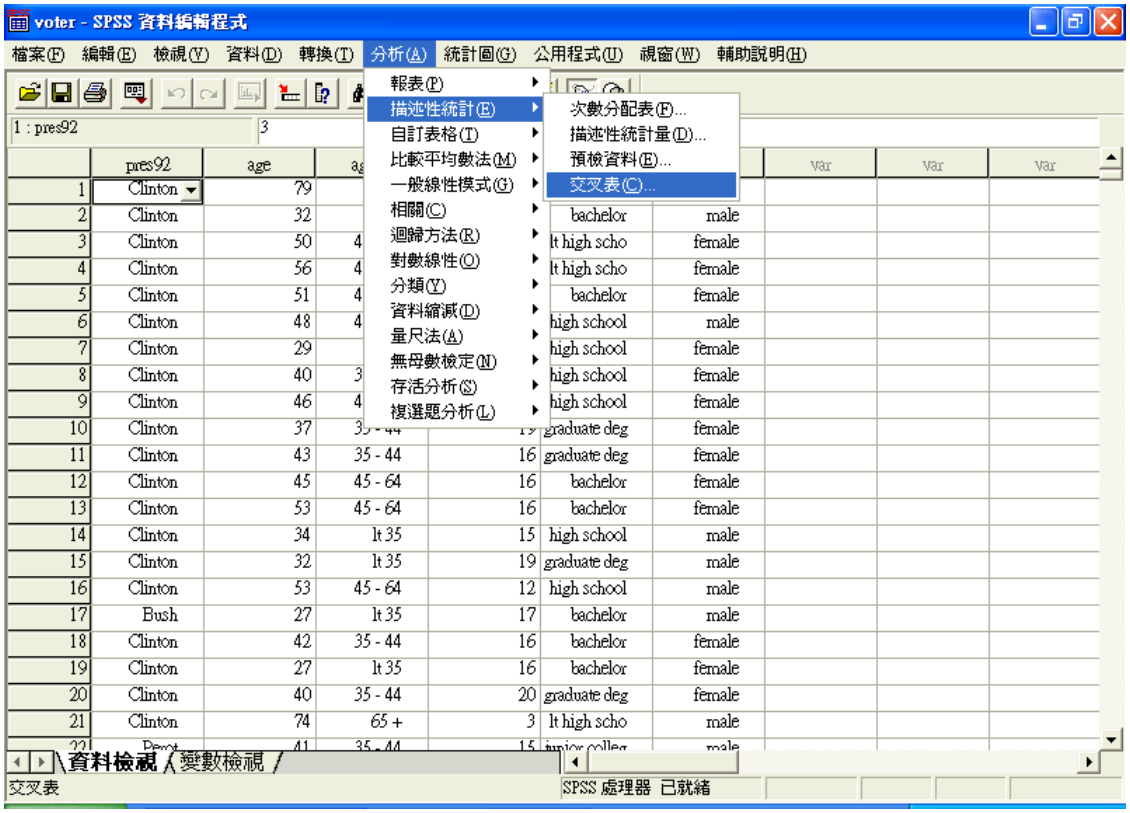
根據上表資料計算可得

$$\sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 33.830 > \chi^2_{0.05} = 5.99$$

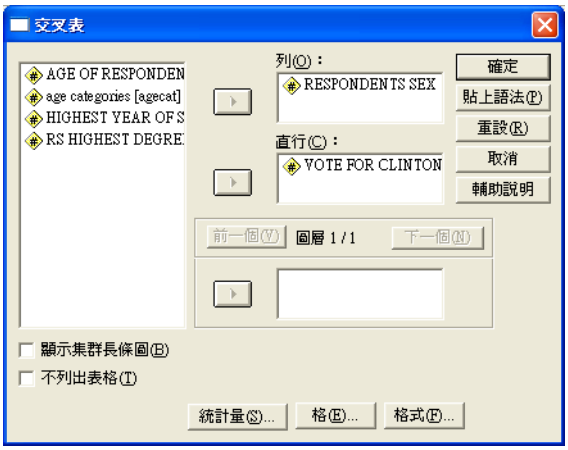
故棄卻虛無假設，即選民性別與投票傾向有關。

SPSS之操作步驟如下：

Step 1. SPSS功能列⇒分析⇒描述性統計⇒交叉表



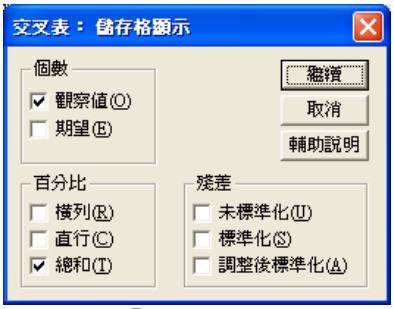
Step 2.



Step 3. 統計量選項中選擇『卡方統計量』；格選項中選擇顯示『期望』及「總和』百分比



(a). 卡方統計量



(b). 『期望』及「總和』

由SPSS所輸出之卡方檢定表可知，卡方統計量爲33.830。此時P-value = 0.000，故棄卻虛無假設，即選民性別與投票傾向是否有關。

卡方檢定

	數值	自由度	漸近顯著性 (雙尾)
Pearson卡方	33.830	2	0.000
概似比	33.866	2	0.000
線性對線性的關連	19.360	1	0.000
有效觀察值的個數	1847		

3.3 齊一性檢定

齊一性檢定爲檢定各組獨立樣本，是否來自同一母體或母體分配是否一致？例如：台北市選民的政黨比例是否與高雄市選民的政黨比例一致。假設有$A_1, \ldots, A_h$等h個母體，將這些母體以某一共同標準分割成$B_1, \ldots, B_k$組。

		B				
		B_1	B_2	$\cdots$	B_k	總計
	A_1	O_{11}	O_{12}	$\cdots$	O_{1c}	n_1
A	A_2	O_{21}	O_{22}	$\cdots$	O_{2c}	n_2
		$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
	A_h	O_{h1}	O_{h2}	$\cdots$	O_{hk}	n_k
	總計	O_1	O_2	$\cdots$	O_k	$O = N$

上表可改以機率形式表示爲

		B				
		B_1	B_2	$\cdots$	B_k	總計
	A_1	P_{11}	P_{12}	$\cdots$	P_{1c}	1
A	A_2	P_{21}	P_{22}	$\cdots$	P_{2c}	1
		$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
	A_h	P_{r1}	P_{r2}	$\cdots$	P_{rc}	1

本檢定以檢定各組獨立樣本是否來自同一母體或母體分配是否一致，其統計假設爲

$$\begin{cases} H_0 : P_{1j} = P_{2j} = \cdots = P_{hj}, j = 1, 2, \ldots, k \\ H_1 : P_{ij}\text{不全相等}, \ j = 1, 2, \ldots, k \end{cases}$$

各組之觀察值O_{ij}所相對應之理論次數爲$E_{ij} = n_i \times P_{ij}$。若虛無假設正確，

$$P_{1j} = P_{2j} = \cdots = P_{hj} = P_j = \frac{O_j}{N}$$

所以

$$E_{ij} = n_i \times P_{ij} = \frac{n_i \times O_j}{N}$$

因此，期望次數爲行總和與列總和的乘積再除以樣本大小，即

$$\sum_{i=1}^{h} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2 \left((r-1)(c-1) \right) \tag{3-9}$$

由上式可知，當H_0不爲眞時，檢定統計量會變大，所以當檢定統計量落入棄卻域

$$C = \left\{ \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} > \chi_\alpha \left((r-1)(c-1) \right) \right\} \tag{3-10}$$

時，則應棄卻虛無假設H_0。

例 3-4. 以SPSS內部範例檔voter.sav爲例，試在$\alpha = 0.05$下，檢定1992年美國總統大選選民性別與投票傾向是否有關？

$$\begin{cases} H_0：\text{不同學歷之選民投票行爲一致} \\ H_1：\text{不同學歷之選民投票行爲不一致} \end{cases}$$

由SPSS報表之結果輸出如下：

卡方檢定

	數值	自由度	漸近顯著性 (雙尾)
Pearson卡方	35.516	8	0.000
概似比	37.703	8	0.000
線性對線性的關連	0.217	1	0.642
有效觀察值的個數	1847		

由卡方檢定表可知，卡方統計量爲36.153，此時P-value $= 0.000$。因此，棄卻虛無假設，即不同學歷之選民投票行爲不一致。

第 4 章

變異數分析

統計資料常受多種因素的影響，而使各個體的某種特徵發生變異，實驗者經常會要比較超過兩個處理的問題，例如：多個不同品種小麥的平均產量是否相同，多種不同廠牌的汽車平均每加侖可以行駛的里程數是否不同。而對這種影響因素所造成之變異的觀察與驗證的統計方法，稱爲變異數分析。

4.1 名詞解釋及基本假設

4.1.1 名詞解釋

1 因子（Factor）：調查中研究某一獨立變數或爲引起資料發生變動的原因。例如： 價格對於銷售量之影響，價格即爲一因子。同樣地，研究四種不同電視節目對觀衆吸引力之影響，其中電視節目即爲一因子。

2 因子水準（Levels of Factor）：因子之特殊形式（所包含的數目）或爲表示因子狀態之條件。例如上例中之價格（X =50元，60元，70元）共有三個水準。

3 **一因子分類**：研究之對象只包含一個獨立變數或事項者。意即觀察值以一個標準爲分類基礎稱之。

4 **多因子分類**：研究之對象只包含兩個或以上獨立變數或事項者。意即觀察值以多個標準爲分類基礎稱之。

5 **屬質因子**：因子水準以屬質之方式來表示稱之，如廣告型態（電視，報紙，雜誌廣告）。

6 **屬量因子**：因子水準可以用計量表示者，如價格（50元，60元，70元），溫度反應（50°，60°，70°）

7 **處理 (** Treatment **)**：施行於實驗單位之特定的實驗條件。

8 **完全隨機化設計 (** Completely Randomized Design **)** ：所有的實驗單位，皆以隨機指定到所有的處理水準的實驗設計。

9 **區集設計 (** Block Design **)** ：，特性相同或類似的的實驗單位組成區集，在各區集內隨機安排實驗單位接受指定的處理。

4.1.2 假設

1 每一因子水準所對應之機率分配皆服從常態分配。

2 所有樣本都是隨機抽取而得，且彼此獨立。

3 各常態母體之變異數皆相等。

4.2 一因子變異數分析

一因子變異數分析係研究之對象只包含一個獨立變數，而觀察此因素之不同對研究對象的影響是否有顯著差異，其數學模式如下：

$$X_{ij} = \mu_i + \varepsilon_{ij},\ \varepsilon_{ij} \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right) \tag{4-1}$$

其中 ，

X_{ij} 第i個因子水準之第j的樣本觀測值

μ_i 第i個母體之平均數

ε_{ij} 隨機誤差項

假設實驗者由k個獨立常態母體中抽出k組隨機樣本，資料如下：

母體	U_1	U_2	$\cdots$	U_i	$\cdots$	U_k	
樣本	X_{11}	X_{21}	$\cdots$	X_{i1}	$\cdots$	X_{k1}	
	X_{12}	X_{22}	$\cdots$	X_{i2}		X_{k2}	
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	
	X_{1n_1}	X_{2n_2}	$\cdots$	X_{in_i}	$\cdots$	X_{kn_k}	
總合	$T_{1.}$	$T_{2.}$	$\cdots$	$T_{i.}$	$\cdots$	$T_{k.}$	$T_{..}$
平均	$\overline{X}_{1.}$	$\overline{X}_{2.}$	$\cdots$	$\overline{X}_{i.}$	$\cdots$	$\overline{X}_{k.}$	$\overline{X}_{..}$

首先將$N(=n_1+n_2+\cdots+n_k)$個樣本所求得之總變異分割得

$$\begin{aligned}\text{SST} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\overline{X}_{..}\right)^2 \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\overline{X}_{i.}\right)^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\overline{X}_{i.}-\overline{X}_{..}\right)^2 \\ &= \text{SSE}+\text{SSB} \end{aligned} \tag{4-2}$$

由以上公式可以發現該組隨機樣本之N個樣本觀測值爲何不全相等，而是有所差異？一部分的原因來自於同組內所造成的變異（SSE，例如：同一個授課教師，課堂上學生的學習成果並非一致），另一部分則是由於組別不同所造成變異（SSB，例如：不同的授課方式所造成學生的學習成果差異）。故若統計假設爲

$$\begin{cases} H_0:\mu_1=\mu_2=\cdots=\mu_k=\mu \\ H_1:\mu_i \text{ 不全相等} \end{cases}$$

當H_0爲眞時，由於$\text{SSB}/\sigma^2\sim\chi^2(k-1)$及$\text{SSE}/\sigma^2\sim\chi^2(N-k)$，可將變異來源可編製成下表：

ANOVA Table

變異來源	平方和	自由度	均方和	F值
組間	SSB	$k-1$	$\text{MSB}=\frac{\text{SSB}}{k-1}$	$F=\frac{\text{MSB}}{\text{MSE}}$
組內	SSE	$N-k$	$\text{MSE}=\frac{\text{SSE}}{N-k}$	
總和	SST	$N-1$		

其中，

$$F=\frac{\text{MSB}}{\text{MSE}}\sim\mathcal{F}\left(k-1,N-k\right)$$

因此，若H_0爲眞，則表示樣本觀測值間之所以有所差異，其主要的原因來自於組內所造成，故$F=\text{MSB}/\text{MSE}$必然很小；反之若H_0不爲眞，即μ_i不全相等，則表示樣本觀測值間之所以有所差異，其主要的原因來自於組別不同所造成。此時，$F=\text{MSB}/\text{MSE}$必然很大，故檢定之棄卻域爲

$$C=\left\{F=\frac{\text{MSB}}{\text{MSE}}>F_\alpha\left(k-1,N-k\right)\right\} \tag{4-3}$$

例 4-1. Montgomery認爲人造纖維的強度可能會受纖維中含棉的比率所影響。在考慮的五種比率值中每一種皆取五個觀測值如下，在顯著水準下，利用$\mathcal{F}$檢定來看由於使用含棉的比率不同而拉張強度是否有所差異。

含棉比率	拉張強度				
15	7	7	15	11	9
20	12	17	12	18	18
25	14	18	18	19	19
30	19	25	22	19	23
35	7	10	11	15	22

解：經計算可求得變異來源如下：

ANOVA Table

變異來源	平方和	自由度	均方和	F值
組間	401.84	4	100.46	7.657
組內	262.40	20	13.12	
總和	664.24	24		

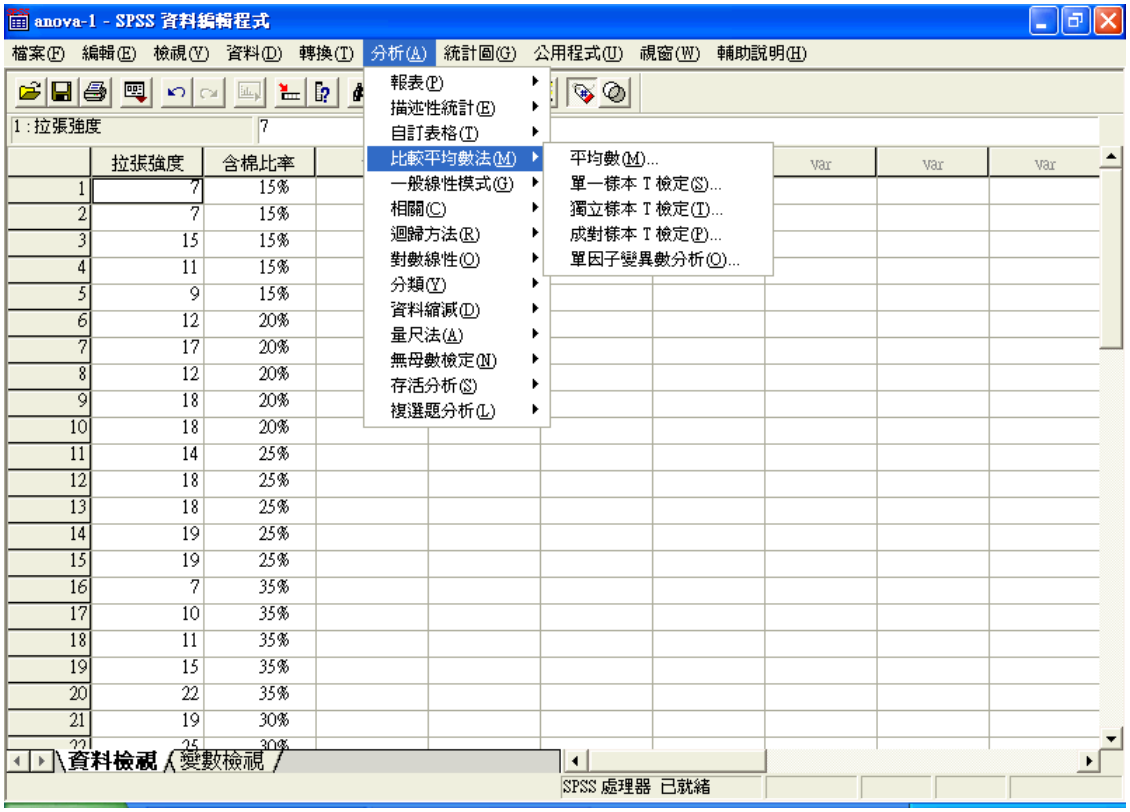
因爲F $= 7.657 > F_{0.05}(4, 20) = 2.8661$，所以棄卻虛無假設

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 \\ H_1 : \mu_i \text{ 不全相等} \end{cases}$$

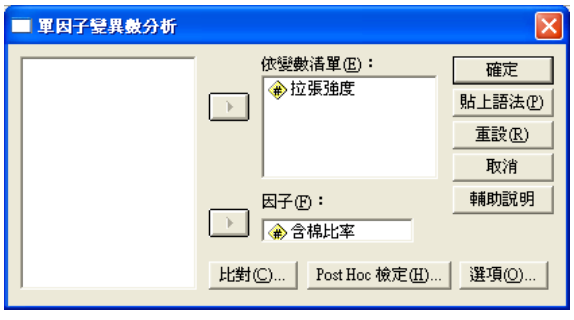
即含棉比率不同，人造纖維之拉張強度也會有所不同。

SPSS之操作步驟如下：

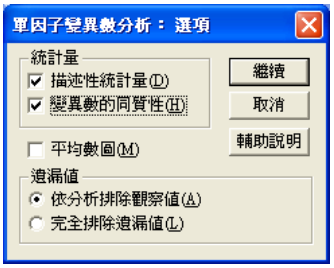
Step 1. SPSS功能列⇒分析⇒比較平均數法⇒單因子變異數分析



Step 2. 選擇變數與因子



Step 3. 在選項中選取『敍述性統計量』及『變異數的同質性』



由『變異數同質性檢定』表可知，因爲P-value $= 0.149 > 0.05$，故不棄卻虛無假設。

變異數同質性檢定

拉張強度

Levene 統計量	分子自由度	分母自由度	顯著性
1.905	4	20	0.149

又在『變異數分析表中』之P-value $= 0.001 < 0.05$，因此棄缺虛無假設，即人造纖維之拉張強度也會有所不同。

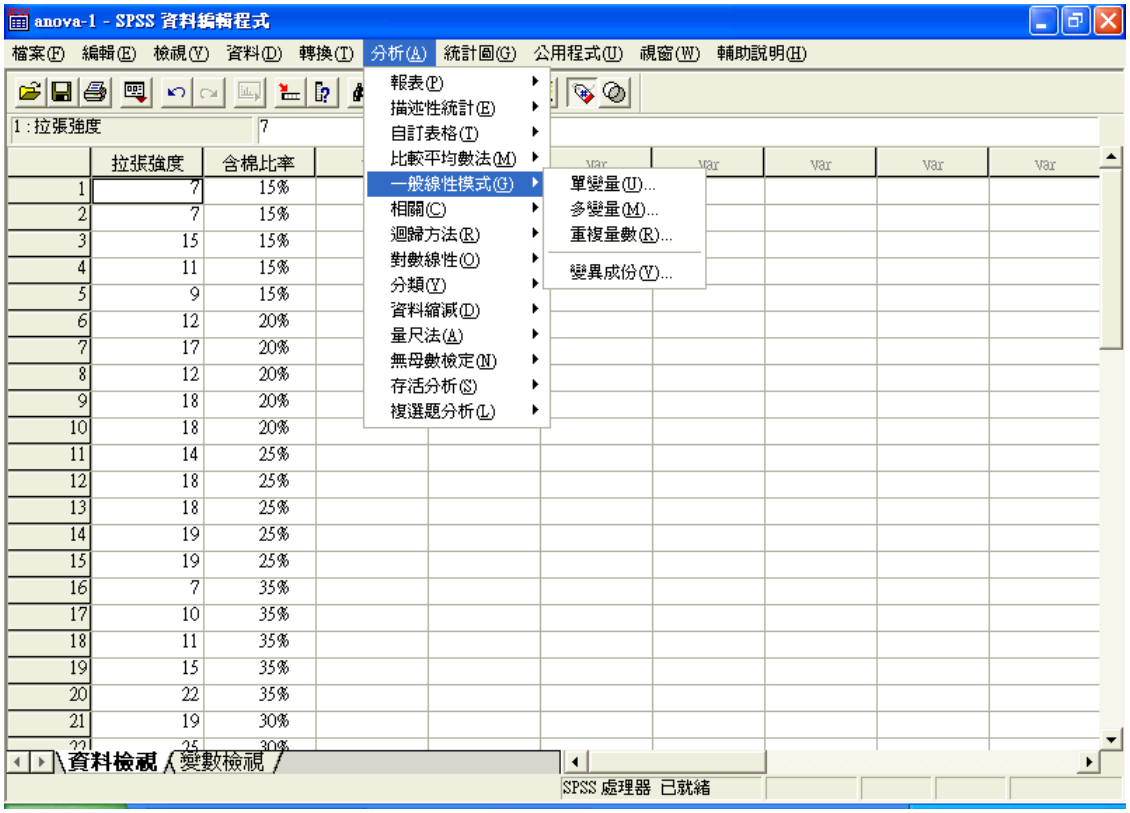
變異數分析

拉張強度

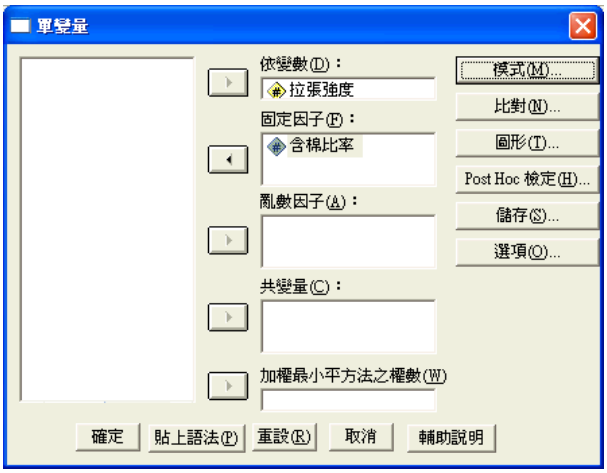
	平方和	自由度	平均平方和	F 檢定	顯著性
組間	401.840	4	100.460	7.657	0.001
組內	262.400	20	13.120		
總和	664.240	24			

除了單因子變異數分析方法外，SPSS中亦可由一般線性模式來解決此一問題，其操作步驟如下：

Step 1. SPSS功能列⇒分析⇒一般線性模式⇒單變量



Step 2. 選擇變數與固定因子



由於SPSS輸出報表中『含棉比率』之P-value $= 0.001 < 0.05$，因此棄缺虛無假設，即人造纖維之拉張強度也會有所不同。

變異數同質性檢定

Levene 統計量	分子自由度	分母自由度	顯著性
0.106	2	394	0.900

變異數分析

	平方和	自由度	平均平方和	F 檢定	顯著性
組間	7984.957	2	3992.479	97.969	0.000
組內	16056.415	394	40.752		
總和	24041.372	396			

受試者間效應項的檢定
拉張強度

來源	型 III 平方和	自由度	平均平方和	F 檢定	顯著性
校正後的模式	401.840	4	100.460	7.657	0.001
截距	5990.760	1	5990.760	456.613	0.000
含棉比率	401.840	4	100.460	7.657	0.001
誤差	262.400	20	13.120		
總和	6655.000	25			
校正後的總數	664.240	24			

4.3 二因子變異數分析

在一因子變異數分析中，因將其他條件固定，而只針對某一因子來解析，依其因子水準不同做隨機化之實驗配置，故其實驗結果比較狹窄，而同時對影響大之原因隨機化，故會使誤差變異加大，導致差異而使得檢定及推定的效率降低。因此要使得檢定效率提高，則必須探討多因子分析。其數學模式如下：

$$X_{ij} = \mu_{ij} + \varepsilon_{ij},\ \varepsilon_{ij} \sim \mathcal{N}\left(0, \sigma^2\right) \tag{4-4}$$

其中，

X_{ij} 表列因子第i個因子水準之第j的樣本觀測值

$\mu_{i.}$ 表列因子第i個母體之平均數

$\mu_{.j}$ 表行因子第j個母體之平均數

ε_{ij} 表隨機誤差項

假設實驗資料如下：

		行因子							
		1	2	$\cdots$	i	$\cdots$	R	總和	平均
列因子	1	X_{11}	X_{21}	$\cdots$	X_{i1}	$\cdots$	X_{R1}	$T_{1.}$	$\overline{X}_{1.}$
	2	X_{12}	X_{22}	$\cdots$	X_{i2}		X_{R2}	$T_{2.}$	$\overline{X}_{2.}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$
	C	X_{1C}	X_{2C}	$\cdots$	X_{iC}	$\cdots$	X_{RC}	$T_{R.}$	$\overline{X}_{R.}$
總和		$T_{.1}$	$T_{.2}$	$\cdots$	$T_{.i}$	$\cdots$	$T_{C.}$	$T_{..}$	
平均		$\overline{X}_{.1}$	$\overline{X}_{.2}$	$\cdots$	$\overline{X}_{.i}$	$\cdots$	$\overline{X}_{C.}$		$\overline{X}_{..}$

所以總變異可寫成

$$\begin{aligned} \text{SST} &= \sum_{i=1}^{r}\sum_{j=1}^{c}\left(\overline{X}_{i.}-\overline{X}_{..}\right)^2+\sum_{i=1}^{r}\sum_{j=1}^{c}\left(\overline{X}_{.j}-\overline{X}_{..}\right)^2 \\ &\quad +\sum_{i=1}^{r}\sum_{j=1}^{c}\left(X_{ij}-\overline{X}_{i.}-\overline{X}_{.j}+\overline{X}_{..}\right)^2 \\ &= \text{SSR}+\text{SSC}+\text{SSE} \end{aligned} \tag{4-5}$$

因此，可將ANOVA Table編制如下：

ANOVA Table

變異來源	均方和	自由度	平方和	F值
SSR	$\sum_{i=1}^{r}\sum_{j=1}^{c}\left(\overline{X}_{i.}-\overline{X}_{..}\right)^2$	$r-1$	$\frac{\text{SSR}}{r-1}$	$F_a=\frac{\text{MSR}}{\text{MSE}}$
SSC	$\sum_{i=1}^{r}\sum_{j=1}^{c}\left(\overline{X}_{.j}-\overline{X}_{..}\right)^2$	$c-1$	$\frac{\text{SSC}}{c-1}$	$F_b=\frac{\text{MSC}}{\text{MSE}}$
SSE	$\sum_{i=1}^{r}\sum_{j=1}^{c}\left(X_{ij}-\overline{X}_{i.}-\overline{X}_{.j}+\overline{X}_{..}\right)^2$	$(r-1)(c-1)$	$\frac{\text{SSE}}{(r-1)(c-1)}$	
SST	$\sum_{i=1}^{r}\sum_{j=1}^{c}\left(X_{ij}-\overline{X}_{..}\right)^2$	$rc-1$		

利用ANOVA Table檢定下列假設

1

$$\begin{cases} H_0:\mu_{1.}=\mu_{2.}=\cdots=\mu_{r.} \\ H_1:\mu_{1.},\mu_{2.},\cdots,\mu_{r.}\text{不全相等} \end{cases}$$

棄卻域爲

$$C=\left\{F_a=\frac{\text{MSR}}{\text{MSE}}>F_\alpha\left(r-1,(r-1)(c-1)\right)\right\} \tag{4-6}$$

2

$$\begin{cases} H_0:\mu_{.1}=\mu_{.2}=\cdots=\mu_{.c} \\ H_1:\mu_{.1},\mu_{.2},\cdots,\mu_{.c}\text{不全相等} \end{cases}$$

棄卻域爲

$$C = \left\{ F_b = \frac{\text{MSC}}{\text{MSE}} > F_\alpha \left(c-1, (r-1)(c-1) \right) \right\} \tag{4-7}$$

例 4-2. 設有三種不同品種的稻米，分別使用四種不同的肥料。今隨機選擇面積相等條件相同的12塊田地做實驗，得到其收穫量（以千公斤計）如下表：

		品種		
		甲	乙	丙
肥料	A	8	3	7
	B	10	4	8
	C	6	5	6
	D	8	4	7

取顯著水準$\alpha = 0.05$，試分別檢定

1 不同肥料

2 不同品種

所得到的平均收穫量是否會有顯著差異？

解：經計算可求得變異來源如下:

ANOVA Table

變異來源	平方和	自由度	均方和	F值
列因子	4.6667	3	1.5556	1.2727
行因子	34.6667	2	17.3333	14.1818
殘差	7.3333	6	1.2222	
總和	46.6667	11		

不同品種稻米之平均收穫量是否相等之棄卻域爲

$$C = \left\{ F = \frac{\text{MSC}}{\text{MSE}} > F_{0.05}(2,6) = 5.1432 \right\}$$

因爲$F = 1.2727 < F_{0.05}(3,6) = 4.7571 \notin C$，所以不棄卻$H_0$：$\mu_A = \mu_B = \mu_C = \mu_D$，即稻米平均收穫量不會因爲品種不同而有所差異。

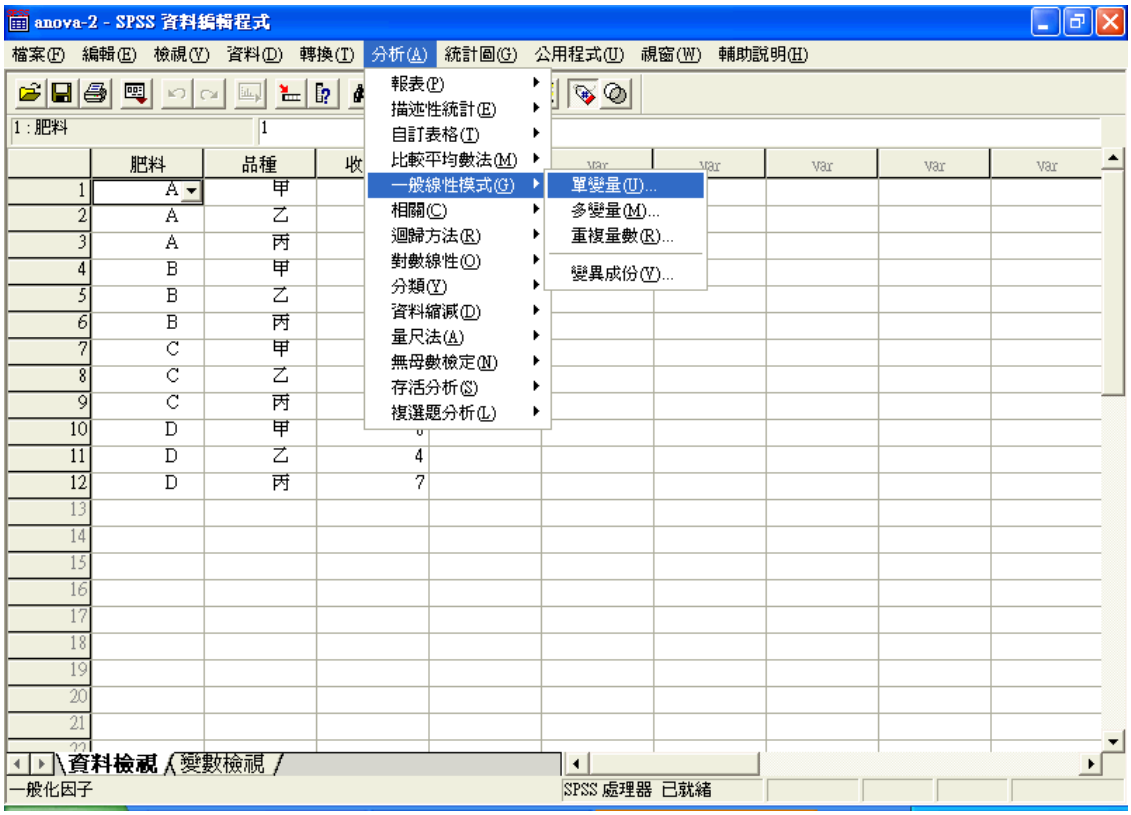
同理，使用不同肥料之平均收穫量是否相等之棄卻域爲

$$C = \left\{ F = \frac{\text{MSR}}{\text{MSE}} > F_{0.05}(3,6) = 4.7571 \right\}$$

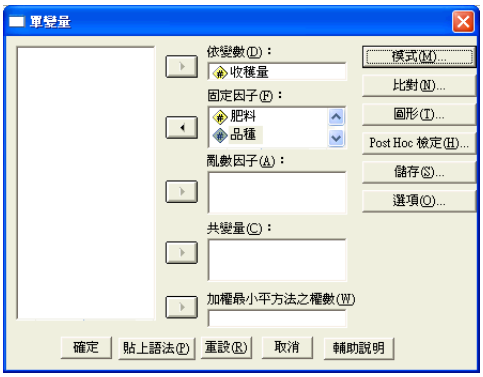
因爲$F = 14.1818 > F_{0.05}(2,6) = 5.1432 \in C$，所以棄卻$H_0$：$\mu_{甲} = \mu_{乙} = \mu_{丙}$ ，即稻米平均收穫量會因爲肥料不同而有所差異。

SPSS之操作步驟如下：

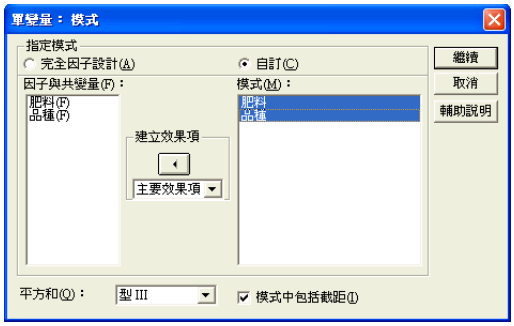
Step 1. SPSS功能列⇒分析⇒一般線性模式⇒單變量



Step 2. 選擇依變數與固定因子



Step 3. 在模式選項中選擇『自訂』，並將肥料及品種兩因子依序由左邊方塊移至右邊



由於SPSS輸出報表中『肥料』之P-value $= 0.365 > 0.05$及『品種』之P-value $= 0.005 < 0.05$，故可知肥料對稻米收穫量無顯著影響；品種則對稻米收穫量有顯著影響。

受試者間效應項的檢定

拉張強度

來源	型 III 平方和	自由度	平均平方和	F 檢定	顯著性
校正後的模式	39.333	5	7.867	6.436	0.021
截距	481.333	1	481.333	393.818	0.000
肥料	4.667	3	1.556	1.273	0.365
品種	34.667	2	17.333	14.182	0.005
誤差	7.333	6	1.222		
總和	528.000	12			
校正後的總數	46.667	11			

4.4 多重比較

當變異數分析F值達顯著水準，表示至少有兩組平均數之間有顯著差異存在，在這許多平均數之間的差異比較之統計檢定方法，就叫做「多重比較檢定」。當然，若F檢定的結果爲不棄卻虛無假設，就沒有太多後續問題；但如果棄卻虛無假設，那麼是哪些比較小，哪些比較大呢？『多重比較』主要的是考慮這類問題，一般都當作ANOVA之後的輔助問題來看。文獻上多重比較的方法至少十種以上，一般常用的有

1 Fisher LSD法：簡單易懂、容易計算；但當組數增大時，型I誤差發生機率變大。

$$\overline{X}_i - \overline{X}_j \pm t_{\frac{\alpha}{2}} (N-k) \sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \tag{4-8}$$

2 Bonferroni法：改良自LSD法，適用於比對個數較少時。

$$\overline{X}_i - \overline{X}_j \pm t_{\frac{\alpha}{2\binom{k}{2}}} (N-k) \sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \tag{4-9}$$

3 Scheffé法：適用於比對個數較多時，且結論較爲保守。

$$\overline{X}_i - \overline{X}_j \pm \sqrt{(k-1) F_\alpha (k-1, N-k)} \sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \tag{4-10}$$

4 Tukey法：適用於各組樣本大小相等的場合。若各組樣本大小不等，則用Tukey-Kramer法較爲適宜。

$$\overline{X}_i - \overline{X}_j \pm \frac{1}{\sqrt{2}} q_\alpha (k, N-k) \sqrt{\frac{\text{MSE}}{n}} \tag{4-11}$$

以上各種方法之主要差異，在於求聯立信賴區間時使用不同的機率理論，並無特別優劣之分。以例題4-1爲例，SPSS對含棉比率做事後比較之操作爲在『Post Hoc』選項中勾選所欲進行比較方法。以Bonferroni法說明如下：

多重比較

拉張強度
Bonferroni 法

(I) 含棉比率	(J) 含棉比率	平均差異 (I-J)	標準誤	顯著性	95% 信賴區間	
					下界	上界
15%	20%	−5.60	2.29	0.239	−12.82	1.62
	25%	−7.80	2.29	0.028	−15.02	−0.58
	30%	−11.80	2.29	0.000	−19.02	−4.58
	35%	−3.20	2.29	1.000	−10.42	4.02
20%	15%	5.60	2.29	0.239	−1.62	12.82
	25%	−2.20	2.29	1.000	−9.42	5.02
	30%	−6.20	2.29	0.136	−13.42	1.02
	35%	2.40	2.29	1.000	−4.82	9.62
25%	15%	7.80	2.29	0.028	0.58	15.02
	20%	2.20	2.29	1.000	−5.02	9.42
	30%	−4.00	2.29	0.961	−11.22	3.22
	35%	4.60	2.29	0.583	−2.62	11.82
30%	15%	11.80	2.29	0.000	4.58	19.02
	20%	6.20	2.29	0.136	−1.02	13.42
	25%	4.00	2.29	0.961	−3.22	11.22
	35%	8.60	2.29	0.012	1.38	15.82
35%	15%	3.20	2.29	1.000	−4.02	10.42
	20%	−2.40	2.29	1.000	−9.62	4.82
	25%	−4.60	2.29	0.583	−11.82	2.62
	30%	−8.60	2.29	0.012	−15.82	−1.38

由上表可發現，25%及30%含棉比率所生產出來的人造纖維的強度顯著比15%含棉比率所生產出來的人造纖維來得高，但與20%含棉比率所生產出來的人造纖維相比較，則無顯著差異。另外，35%含棉比率所生產出來的人造纖維的強度顯著比30%含棉比率所生產出來的人造纖維來得低，但與其他含棉比率所生產出來的人造纖維相比較，則無顯著差異。

4.5 殘差分析

由於以上分析均在$\varepsilon_{ij} \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right)$假設下進行。若其偏離假設太遠，則此變異數分析模型就有問題，所以有必要去檢視這些假設是否成立。這些假設的檢視稱爲殘差分析，在變異數模式中，殘差e_{ij}之定義爲

$$e_{ij} = X_{ij} - \overline{X}_{i.}$$

殘差分析大多以圖形檢查爲基礎，檢定的方向有三：

1 常態性：通常以檢查殘差之常態機率圖爲主，所繪出之點越靠近45度線，支持殘差爲常態分配的假設證據就越強。若以檢定方式，則可以卡方適合度檢定或 K-S 檢定（ Kolmogorov-Smirnov goodness-of-fit test ） 進行。當資料不服從常態分配，怎可改採無母數統計分析中之 Kruskal-Wallis test （獨立樣本）或 Friedman test （相依樣本）來進行母體平均數之比較。最後必須注意的是當樣

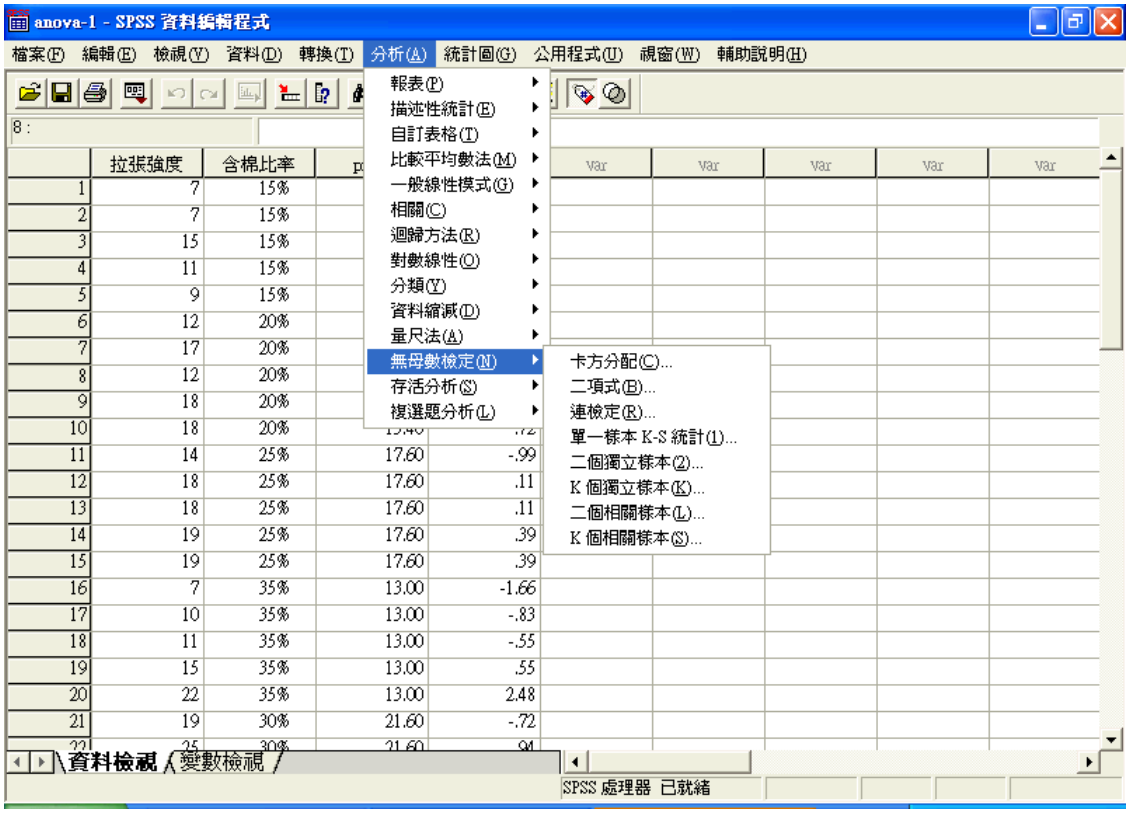
本數小時，K-S檢定的結果較不容易棄卻常態性假說。而無母數統計分析方法在樣本數小做檢定結果的檢定例通常也較小，故也較無實用價值。

2 同質性：以殘差對各組平均數圖檢視同質性，或以檢定方式檢定各母體之變異數是否相等。在SPSS中以Levene's test檢定之。

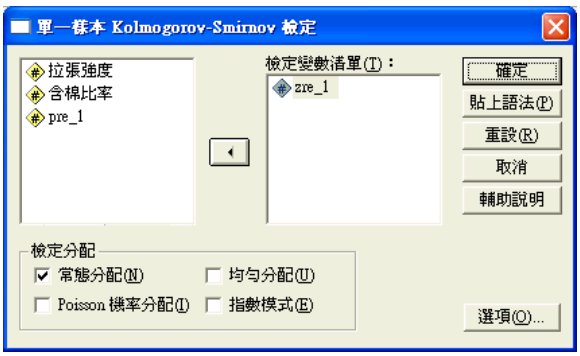
3 獨立性：依收集資料的先後順序，繪殘差點圖。若資料是彼此獨立時，殘差應會隨機散佈在圖上，換言之，殘差應不會成群出現在零線（即原點）的某一方，否則表示非獨立。

以例題4-1爲例，讀者可在一般線性模式單變量分析『儲存』中勾選『標準化殘差』以及『未標準化預測值』以取得資料。首先，在常態性方面，由Q-Q圖發現觀測值接於對角線，因此可推斷殘差應爲常態分配。若欲以嚴謹檢定方式，則可由以下步驟進行K-S檢定：

Step 1. SPSS功能列⇒分析⇒無母數檢定⇒單一樣本K-S統計



Step 2. 選擇欲檢定之變數



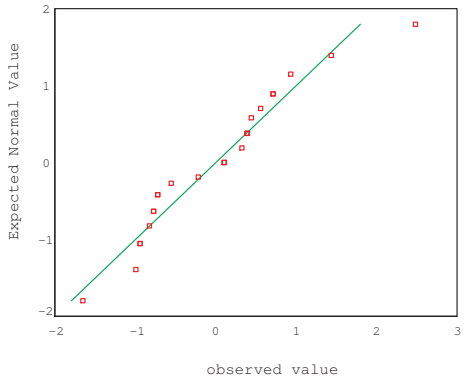
由SPSS所輸出之報表可知，P-value $= 0.676 > 0.05$，因此不棄卻虛無假設

$$\begin{cases} H_0：殘差服從常態分配 \\ H_1：殘差不服從常態分配 \end{cases}$$

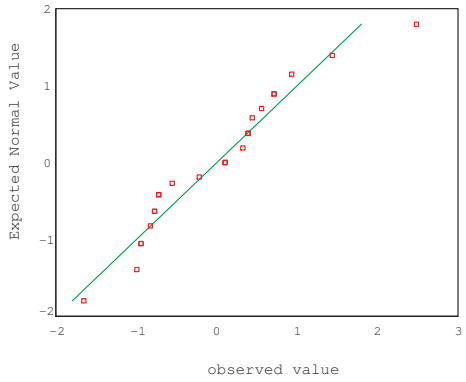
即殘差應服從常態分配。而同質性方面，可由例4-1之變異數同質性檢定報表中發現P-value $= 0.149 > 0.05$，因此不棄卻虛無假設，即各組之變異數應爲相等。最後，由標準化殘差之散佈圖可發現殘差隨機散佈在圖上，資料應是彼此獨立。

單一樣本 Kolmogorov-Smirnov 檢定

		ZRE_1
個數		25
常態參數	平均數	0.0000
	標準差	0.9129
最大差異	絕對	0.144
	正的	0.144
	負的	−0.108
Kolmogorov-Smirnov Z 檢定		0.721
漸近顯著性 (雙尾)		0.676



(a) Q-Q圖



(b) 殘差圖

第 5 章

簡單線性迴歸分析

日常生活中，有許多現象之間都有著相互的關聯。例如，身高與體重或父母與子女的身高。在這些關係中，它們之間關聯的程度不同，密切性也不同。甚至有些關係，我們只能夠說兩者之間相互影響，但究竟如何影響、誰是因誰是果，甚至於沒有因果而只是簡單的伴隨關係，這種時候，我們就可以使用相關與迴歸分析來研究和解釋兩個變量之間的相互關係。

本章僅探討簡單線性迴歸模式，即利用一個解釋變數（X）對應變數（Y）做預測之模式。所謂模式，乃指找出X與Y之間的函數關係，即$Y = f(X)$。然而由於Y爲一隨機變數，所以我們並不能正確的預測Y值是多少。因此我們僅能估計當$X = x$時，隨機變數Y之期望值爲何，其中實際觀測值與期望值之間的落差即爲估計誤差值，在此以ε表示之。由於我們所要建立之模式爲簡單線性迴歸模式，即找出參數β_0與β_1使得

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{5-1}$$

其中ε_i，$i = 1, 2, \ldots n$爲相互獨立且服從$\mathcal{N}(0, \sigma^2)$。由上式可知，在給定$X_i = x_i$時，Y_i之條件期望值及條件變異數分別爲

$$\mathrm{E}\left[Y_i | X_i = x_i\right] = \beta_0 + \beta_1 x_i, i = 1, 2, \ldots, n$$

及

$$\mathrm{Var}\left[Y_i | X_i = x_i\right] = \sigma^2, i = 1, 2, \ldots, n$$

5.1 最小平方法

由於我們所要建立的模式爲

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i\text{，}i = 1, 2, \ldots, n,$$

其中β_0，β_1及σ^2均爲未知參數，通常以最小平方法作爲估計的方法。最小平方法之概念爲找出一組$\widehat{\beta}_0$及$\widehat{\beta}_1$使得誤差平方和爲最小，而且不受ε_i分配限制，一般此種方法普通最小平方法（Method of Ordinarily Least Square），簡稱OLS。 由圖9.1可知，誤差平方和爲$\sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}(Y_i-\beta_0-\beta_1X_i)^2$。由

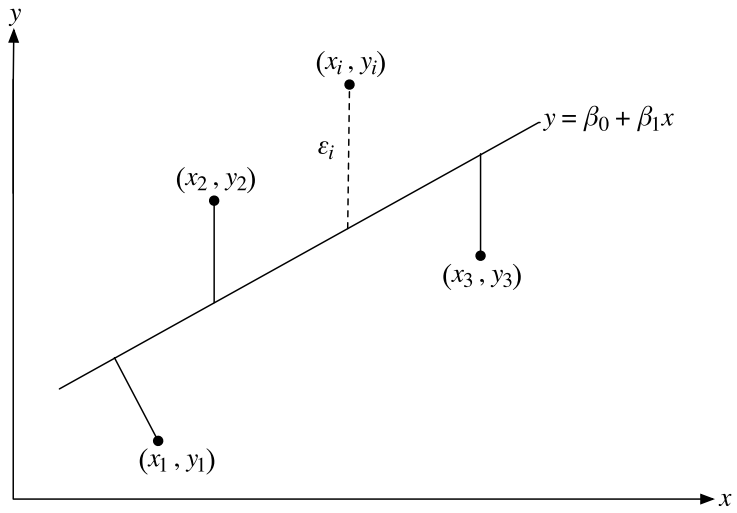


圖 5.1: 普通最小平方法概念圖

於σ^2，β_0，β_1均未知，因此分別以e_i，$\widehat{\beta}_0$及$\widehat{\beta}_1$作爲個別之估計式。所以目標函數可寫成

$$\min \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}\left(Y_i-\widehat{\beta}_0-\widehat{\beta}_1X_i\right)^2 \tag{5-2}$$

分別對$\widehat{\beta}_0$及$\widehat{\beta}_1$做微分，可得最小化之必要條件爲

$$\frac{\partial\sum_{i=1}^{n}e_i^2}{\partial\widehat{\beta}_0} = \sum_{i=1}^{n}2\left(Y_i-\widehat{\beta}_0-\widehat{\beta}_1X_i\right)\times(-1)=0$$

和

$$\frac{\partial\sum_{i=1}^{n}e_i^2}{\partial\widehat{\beta}_1} = \sum_{i=1}^{n}2\left(Y_i-\widehat{\beta}_0-\widehat{\beta}_1X_i\right)\times(-X_i)=0$$

求解二元一次方程組可得

$$\widehat{\beta}_0 = \frac{\sum_{i=1}^{n}Y_i\sum_{i=1}^{n}X_i^2-\sum_{i=1}^{n}X_iY_i\sum_{i=1}^{n}X_i}{n\sum_{i=1}^{n}X_i^2-\left(\sum_{i=1}^{n}X_i\right)^2}$$

及

$$\widehat{\beta}_1 = \frac{n\sum_{i=1}^n X_iY_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n\sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i\right)^2}$$

其中，$\widehat{\beta}_1$及$\widehat{\beta}_0$亦可寫成

$$\begin{aligned}\widehat{\beta}_1 &= \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} \\ &= \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2 \sum_{i=1}^n \left(Y_i - \overline{Y}\right)^2}} \frac{\sqrt{\sum_{i=1}^n \left(Y_i - \overline{Y}\right)^2}}{\sqrt{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}}\end{aligned} \tag{5-3}$$

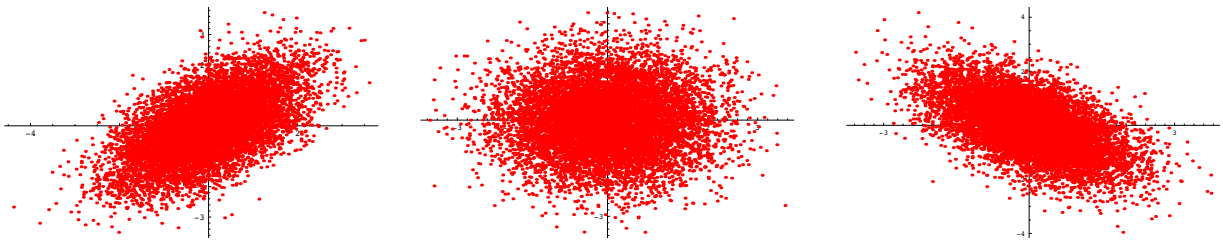
及

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X} \tag{5-4}$$

又樣本相關係數定義爲

$$r_{xy} = \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2 \sum_{i=1}^n \left(Y_i - \overline{Y}\right)^2}} = \frac{\mathrm{Cov}(X,Y)}{S_x}. \tag{5-5}$$

其主要目的爲衡量兩變數線性關係強弱，$-1 \le r \le 1$。當$|r|$越接近1，則表示線性關係越強；反之，當$|r|$越接近0，則表示線性關係越弱。當$r = 0$時，即指這兩變數之間不相關。正相關時，r值在0和1之間，散佈圖是斜向上的，這時一個變數增加，另一個變數也增加；負相關時，r值在-1和0之間，散佈圖是斜向下的，此時一個變數增加，另一 個變數將減少。所以估計式$\widehat{\beta}_1$可改寫成$\widehat{\beta}_1 = r_{xy} \times \frac{S_y}{S_x}$。



(a) 正相關　(b) 不相關　(c) 負相關

圖 5.2: 相關係數圖

此外，我們知道反應變數Y有部份可由解釋變數X來解釋，所以可將總變異來源SST分解可如下：

$$\mathrm{SST} = \sum_{i=1}^n \left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^n \left(Y_i - \widehat{Y}_i + \widehat{Y}_i - \overline{Y}\right)^2$$

$$\begin{aligned} &= \sum_{i=1}^{n}\left(Y_i-\widehat{Y}\right)^2+\sum_{i=1}^{n}\left(\widehat{Y}-\overline{Y}\right)^2 \\ &= \text{SSE}+\text{SSR} \end{aligned} \tag{5-6}$$

其中SSR爲可解釋變異，SSE爲不可解釋變異。由於SSE屬眞正的殘差，我們可以利用它來估計母體變異數，它與母體變異數σ^2之關係如下：

$$\text{E}\left(\text{MSE}\right)=\text{E}\left(\frac{\text{SSE}}{n-2}\right)=\sigma^2 \tag{5-7}$$

因此，MSE爲母體變異數σ^2之不偏估計式。

由於 $\text{SST}/\sigma^2\sim\chi^2(n-1)$且卡方分配具可加性，又 $\sum_{i=1}^{n}\left(Y_i-\widehat{Y}\right)^2$ 與 $\sum_{i=1}^{n}\left(\widehat{Y}-\overline{Y}\right)^2$ 相獨立，故可知 SSE/σ^2 及 SSR/σ^2亦爲卡方分配。因爲 $\text{E}(\text{SSE}/\sigma^2)=n-2$，故可推得 $\text{SSE}/\sigma^2\sim\chi^2(n-2)$。由卡方分配之可加性可知$\text{SSR}/\sigma^2\sim\chi^2(1)$。最後，將各項變異來源編製成下列表格

ANOVA Table

變異來源	平方和	自由度	均方和	F值
SSR	$\sum_{i=1}^{n}\left(\widehat{Y}_i-\overline{Y}\right)^2$	1	MSR= $\frac{\sum_{i=1}^{n}\left(\widehat{Y}_i-\overline{Y}\right)^2}{1}$	F = $\frac{\text{MSR}}{\text{MSE}}$
SSE	$\sum_{i=1}^{n}\left(Y_i-\widehat{Y}_i\right)^2$	$n-2$	MSE= $\frac{\sum_{i=1}^{n}\left(Y_i-\widehat{Y}_i\right)^2}{n-2}$	
SST	$\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2$	$n-1$		

5.2 β_0與β_1之統計推論

由於$\widehat{\beta}_1\sim\mathcal{N}\left(\beta_1,\sigma^2/\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right)$，$\text{SSE}/\sigma^2\sim\chi^2(n-2)$，及$\sigma^2$未知，因此由$t$分配之定義可知

$$\frac{\frac{\widehat{\beta}_1-\beta_1}{\sqrt{\sigma^2/\sum_{i=1}^{n}(X_i-\overline{x})^2}}}{\sqrt{\frac{\text{SSE}/\sigma^2}{n-2}}}=\frac{\widehat{\beta}_1-\beta_1}{\sqrt{\text{MSE}/\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}}\sim t\left(n-2\right) \tag{5-8}$$

故由可求得β_1之$(1-\alpha)\times100\%$信賴區間爲

$$\left[\widehat{\beta}_1-t_{\frac{\alpha}{2}}\left(n-2\right)\sqrt{\frac{\text{MSE}}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}},\widehat{\beta}_1+t_{\frac{\alpha}{2}}\left(n-2\right)\sqrt{\frac{\text{MSE}}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}}\right] \tag{5-9}$$

當統計假設爲雙尾檢定時，即

$$\begin{cases} H_0:\beta_1=k \\ H_1:\beta_1\neq k \end{cases}$$

其棄卻域爲

$$C=\left\{\widehat{\beta}_1<k-t_{\frac{\alpha}{2}}\left(n-2\right)\sqrt{\frac{\text{MSE}}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)}}\right\}$$

值得注意的是，當$k=0$時

$$\begin{aligned}\frac{\widehat{\beta}_1-0}{\sqrt{\sigma^2/\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}} &= r\times\frac{\sqrt{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}}{\sqrt{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}}\times\frac{1}{\sqrt{\sigma^2/\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}}\\ &= r\times\sqrt{\frac{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}{\sigma^2}}\end{aligned}$$

及

$$\sqrt{\frac{\text{SSE}/\sigma^2}{n-2}}=\sqrt{\frac{\left(1-r^2\right)\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}{\left(n-2\right)\sigma^2}}=\sqrt{\frac{1-r^2}{n-2}}\times\sqrt{\frac{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}{\sigma^2}}$$

所以

$$\frac{\widehat{\beta}_1-0}{\sqrt{\sigma^2/\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}}\Bigg/\sqrt{\frac{\text{SSE}/\sigma^2}{n-2}} \tag{5-10}$$

$$= r\times\sqrt{\frac{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}{\sigma^2}}\times\sqrt{\frac{n-2}{1-r^2}}\times\sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}} \tag{5-11}$$

$$= \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{5-12}$$

故可知統計量

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\sim t\left(n-2\right)$$

經由以上討論，我們可推得下列結論： 檢定β_1是否爲0與檢定相關係數是否爲0所得結論相同，即

$$\left\{\begin{array}{l}H_0:\beta_1=0\\ H_1:\beta_1\neq 0\end{array}\right.\Leftrightarrow\left\{\begin{array}{l}H_0:\rho=0\\ H_1:\rho\neq 0\end{array}\right.$$

故當

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}<-t_{\frac{\alpha}{2}}\left(n-2\right)\text{或}\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}>t_{\frac{\alpha}{2}}\left(n-2\right)$$

則棄卻虛無假設$H_0:\rho=0$。

簡單迴歸研究中是研究獨立變數X與反應變數Y之關係，故很自然的我們會想到一個問題就是X對Y的影響爲何？而變數X對Y之解釋能力貢獻爲何？在前面我們以了解可利用是否爲0的檢定來探討X對Y是否有影響；我們也可以利用更普遍化之方法，變異數分析法（ANOVA），來探討。

由於總變異SST可分解成SSR（可解釋變異）及SSE（不可解釋變異），即

$$\begin{aligned}\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2 &= \text{SSR}+\text{SSE}\\ &= \sum_{i=1}^{n}\left(\widehat{Y}_i-\overline{Y}\right)^2+\sum_{i=1}^{n}\left(Y_i-\widehat{Y}_i\right)^2\end{aligned}$$

由上式可知，獨立變數X對反應變數Y的變異解釋能力爲

$$R^2=\frac{\text{SSR}}{\text{SST}} \tag{5-13}$$

一般將R^2稱爲判定係數（Coefficient of Determination）。此外，我們也可由R^2之定義推導出下列結果

$$\begin{aligned}R^2 &= \frac{\text{SSR}}{\text{SST}}=\frac{\widehat{\beta}_1^2\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}\\ &= \frac{\left[\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)\right]^2}{\left[\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right]^2}\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}\\ &= \left[\frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}}\right]^2\\ &= r_{xy}^2\end{aligned} \tag{5-14}$$

故可知判定係數即爲樣本相關係數的平方。因此，相關係數平方即表示解釋變數能夠解釋應變數能力。

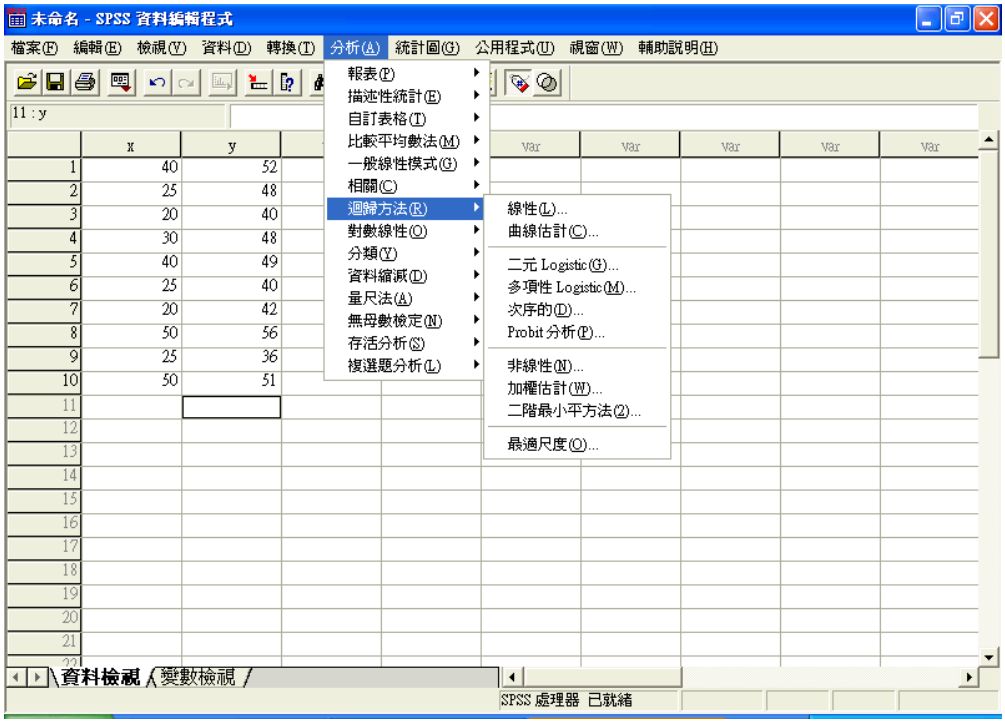
例 5-1. 爲決定廣告費與銷貨額之關係，乃從事一項研究，所得樣本如下：

廣告費：X	40	25	20	30	40	25	20	50	25	50
銷貨額：Y	52	48	40	48	49	40	42	56	36	51

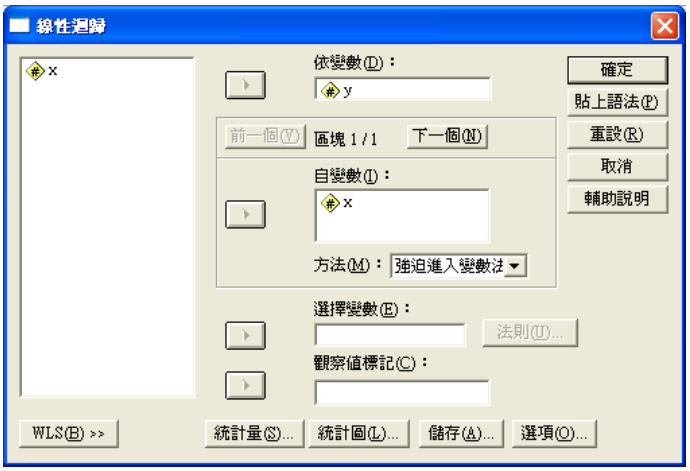
單位：萬元。若假設條件皆符合簡單迴歸模式，試探討廣告費與銷貨額之關係。

解：利用SPSS線性迴歸分析，操作步驟如下：

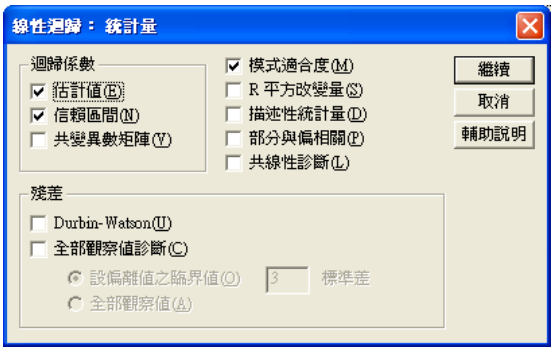
Step 1. SPSS功能列⇒分析⇒迴歸方法⇒線性



Step 2. 選擇依變數與自變數



Step 3. 選取『統計量』對話方塊，並勾選『信賴區間』



SPSS報表輸出如下：首先，由『模式摘要』表可知$R^2 = 0.696$，即銷售量之變異有65.7%來自廣告費。

模式摘要

模式	R	R平方	調過後的R平方	估計的標準誤
1	0.834	0.695	0.657	3.7343

其中R^2算法主要由以下變異數分析表中$\frac{254.041}{365.600}$計算得來。

變異數分析

模式		平方和	自由度	平均平方和	F檢定	顯著性
1	迴歸	254.041	1	254.041	18.218	0.003
	殘差	111.559	8	13.945		
	總和	365.600	9			

又由『係數』之報表可發現，廣告費係數之P-value$= 0.03 < 0.05$。因此，棄卻虛無假設

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

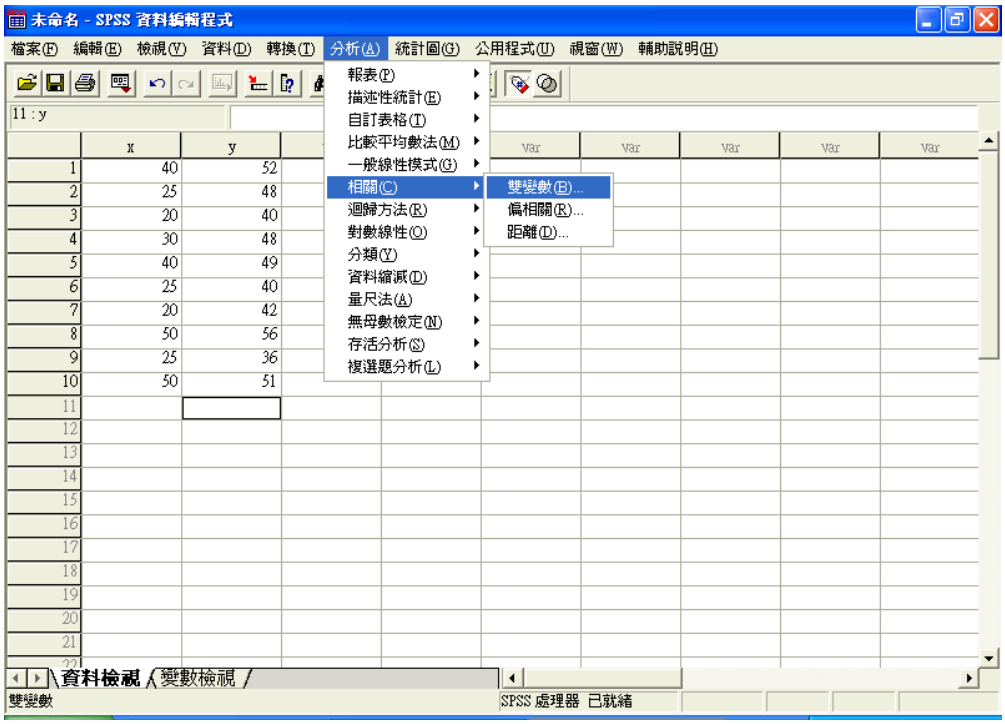
即廣告費支出對銷售量的確會有所影響。

係數

模式		未標準化係數 β之估計值	標準誤	t	顯著性	95% 信賴區間 下限	95% 信賴區間 上限
1	(常數)	31.324	3.680	8.512	0.000	22.838	39.810
	廣告費	0.458	0.107	4.268	0.003	0.210	0.705

此外，也可以相關分析之方式來檢定兩變數之相關係數是否爲0，檢定之方法如下：

Step 1. SPSS功能列⇒分析⇒相關⇒雙變數



Step 2.



SPSS所輸出『相關』報表如下。由於P-value=0.003，所以棄卻虛無假設H_0：$\rho = 0$，即兩變數之相關性顯著存在。

相關

		廣告費	銷售額
廣告費	Pearson 相關	1.000	0.834
	顯著性 (雙尾)	.	0.003
	個數	10	10
銷售額	Pearson 相關	0.834	1.000
	顯著性 (雙尾)	0.003	0.000
	個數	10	10

例 5-2. 以SPSS內部範例檔Employee data.sav爲例，試在$\alpha = 0.05$下，檢定某公司員工學歷高低是否對起新是否有影響?

解：由SPSS所輸出『變異數分析』表可知，P-value $= 0.000 < 0.05$，故棄卻虛無假設，即學歷高低的確對起薪有影響。另外，由邊際檢定的『係數』報表，也可推得相同結論。

變異數分析

模式		平方和	自由度	平均平方和	F檢定	顯著性
1	迴歸	11747808912.32	1	11747808912.32	315.90	0.000
	殘差	17553096053	472	37188762.82		
	總和	29300904965	473			

係數

模式		未標準化係數 β之估計值	標準誤	t	顯著性	95%信賴區間	
						下限	上限
1	(常數)	−6290.97	1340.92	−4.69	0.00	−8925.88	−3656.06
	教育程度	1727.53	97.20	17.77	0.00	1536.54	1918.52

5.3 殘差分析

由於以上分析均在$\varepsilon_i \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right)$假設成立下進行。若其偏離假設太遠，則此迴歸模型就有問題，所以有必要去檢視這些假設是否成立。這些假設的檢視稱爲殘差分析，殘差分析大多以圖形檢查爲基礎，檢定的方向有三：

1 常態性：通常以檢查殘差之常態機率圖爲主，所繪出之點越靠近45度線，支持殘差爲常態分配的假設證據就越強。若以檢定方式，則可以卡方適合度檢定或 K-S 檢定（Kolmogorov-Smirnov goodness-of-fit test）進行。

2 恆常性：繪制殘差e_i與迴歸估計值$\hat{y}_i$及殘差e_i與自變數X的對應圖。若圖型顯示不成一「以零爲中心的水平帶狀」時，則表示殘差不具恆常性；也表示其關係並非線性。此時也可利用將依變數轉換的方式處理（如取對數或開根號等）。

3 獨立性：依收集資料的先後順序，繪殘差點圖。若資料是彼此獨立時，殘差應會隨機散佈在圖上，換言之，殘差應不會成群出現在零線（即原點）的某一方，否則表示非獨立。除了圖示法外，亦可利用Durbin-Watson統計量 來檢定獨立性。

以例5-1爲例，SPSS殘差分析進行步驟如下：

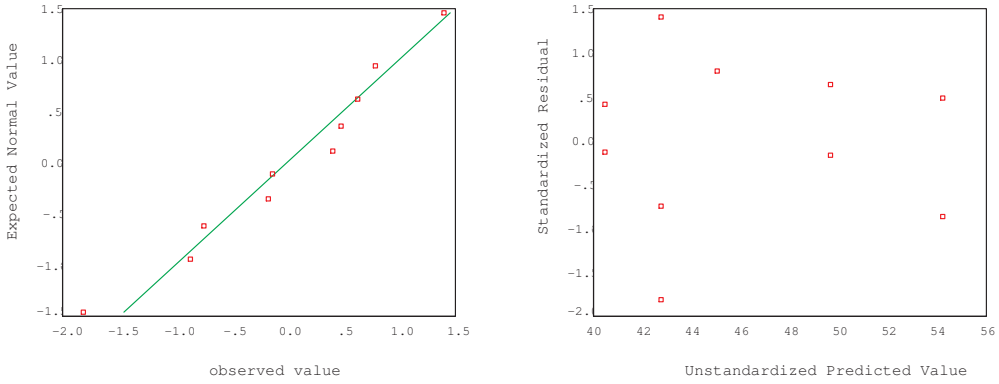
1 取得標準化殘差與估計值，SPSS功能列⇒分析⇒迴歸方法⇒線性⇒選取『儲存』對話方塊，勾選未標準化預測值及標準化殘差，確定後在資料視窗即會增加該兩變數。

2 繪製Q-Q圖，SPSS功能列⇒統計圖⇒Q-Q圖。由圖形可發現，大多數的點都非常接近對角線，故可推斷殘差應服從常態分配。若以K-S檢定，則選擇SPSS功能列⇒分析⇒無母數統計⇒單一樣本K-S檢定。由於P-velue=0.943，所以不棄卻虛無假設H_0：殘差服從常態分配。

單一樣本 Kolmogorov-Smirnov 檢定

		Standardized Residual
個數		10
常態參數	平均數	0.0000000
	標準差	0.9428090
最大差異	絕對	0.167
	正的	0.101
	負的	−0.167
Kolmogorov-Smirnov Z 檢定		0.529
漸近顯著性 (雙尾)		0.943

3 繪制殘差e_i與迴歸估計值$\hat{y}_i$對應圖；SPSS功能列⇒統計圖⇒散佈圖。由圖形可發現，大多數的點均位於以零爲中心的水平帶狀，所以應可推斷殘差具恆常性。又殘差無規則散佈，所以應符合獨立性。



(a) Q-Q圖 (b) 殘差圖

一般而言，若殘差值不服從常態分配時，則需利用變數轉換法將資料轉換成常態資料。常用的方式爲對原始資料取對數、開根號、倒數等。然而，並非所有的的轉換方式皆可將資料轉換成爲常態，在此需檢驗其常態性，直至轉換後的資料服從常態分配。Box和Cox於1964年提出一有系統、有科學根據的冪次轉換方法，其轉換函數定義爲

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \ln y, & \lambda = 0 \end{cases} \tag{5-15}$$

其中，y必須爲正數。該轉換以最大概似函數爲目標函數，求取能使誤差平方和爲最小的次方值（λ），此種次方轉換除了可使資料轉變爲常態分配或近似常態分配外，同時也可使變異之異質性得以改善。 由於Box-Cox轉換所求得之最佳λ值可能爲任意實數值，但實際上常 用的λ值與轉換方式如下：

λ值	轉換方式
1.0	不轉換
0.5	平方根轉換
0.0	對數轉換
-0.5	倒數平方根轉換
-1.0	倒數轉換

在進行資料轉換時，通常會建議使用者使用單純或者近似的λ值來做轉換。舉例來說，當$\lambda = 0.52$時，由於難以解釋，通常取近似的$\lambda = 0.5$，如此一來解釋要容易的多。以下列汽車時速與煞車距離資料爲例，經由迴歸分析所繪得知殘差圖可發現並不符合恆常性之假設，也表示其關係並非線性。因此，需經由變數轉換的方式處理。

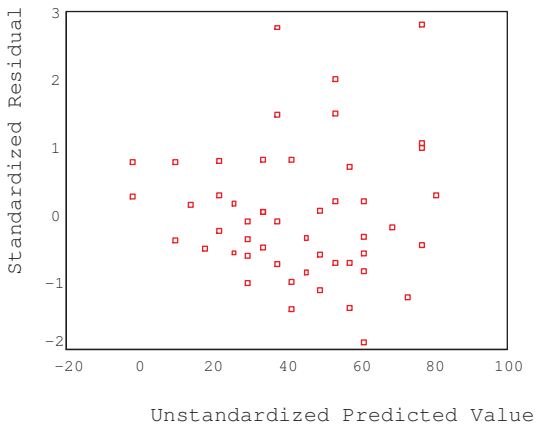
id	速度（X）	煞車距離（Y）	id	速度（X）	煞車距離（Y）
1	4	2	26	15	54
2	4	10	27	16	32
3	7	4	28	16	40
4	7	22	29	17	32
5	8	16	30	17	40
6	9	10	31	17	50
7	10	18	32	18	42
8	10	26	33	18	56
9	10	34	34	18	76
10	11	17	35	18	84
11	11	28	36	19	36
12	12	14	37	19	46
13	12	20	38	19	68
14	12	24	39	20	32
15	12	28	40	20	48
16	13	26	41	20	52
17	13	34	42	20	56
18	13	34	43	20	64
19	13	46	44	22	66
20	14	26	45	23	54
21	14	36	46	24	70
22	14	60	47	24	92
23	14	80	48	24	93
24	15	20	49	24	120
25	15	26	50	25	85

模式摘要

模式	R	R平方	調過後的R平方	估計的標準誤
1	0.807	0.651	0.644	15.38

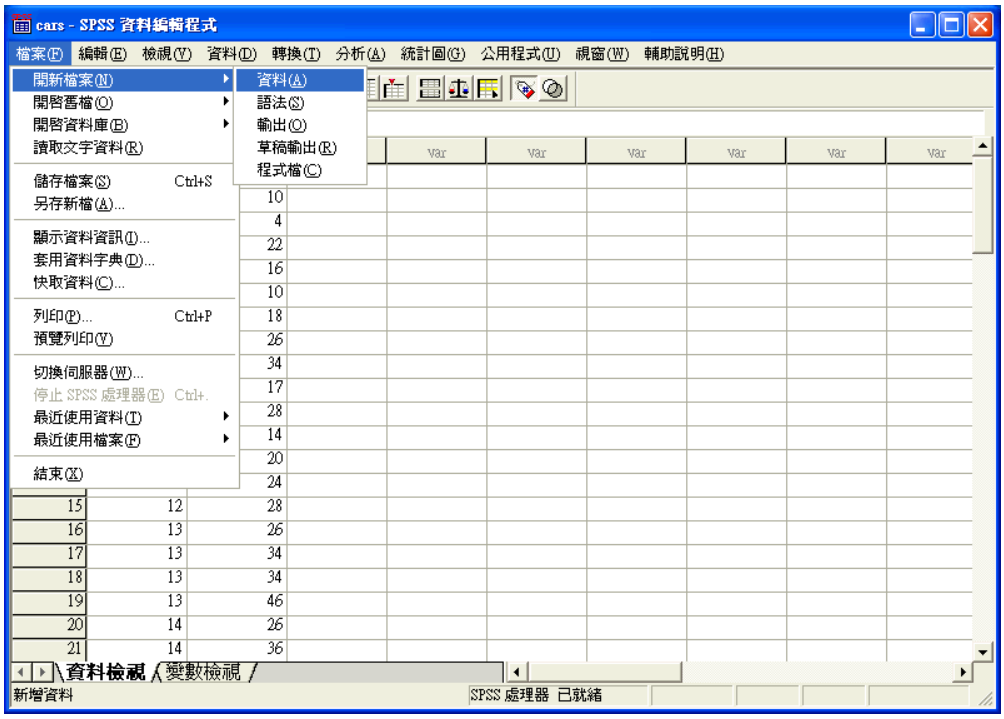
變異數分析

模式		平方和	自由度	平均平方和	F檢定	顯著性
1	迴歸	21185.459	1	21185.460	85.567	.000
	殘差	11353.521	48	263.532		
	總和	32538.980	48			



由於SPSS並無Box-Cox轉換之操作介面，所以必須藉由執行程式語法之功能才可執行。SPSS操作步驟如下：

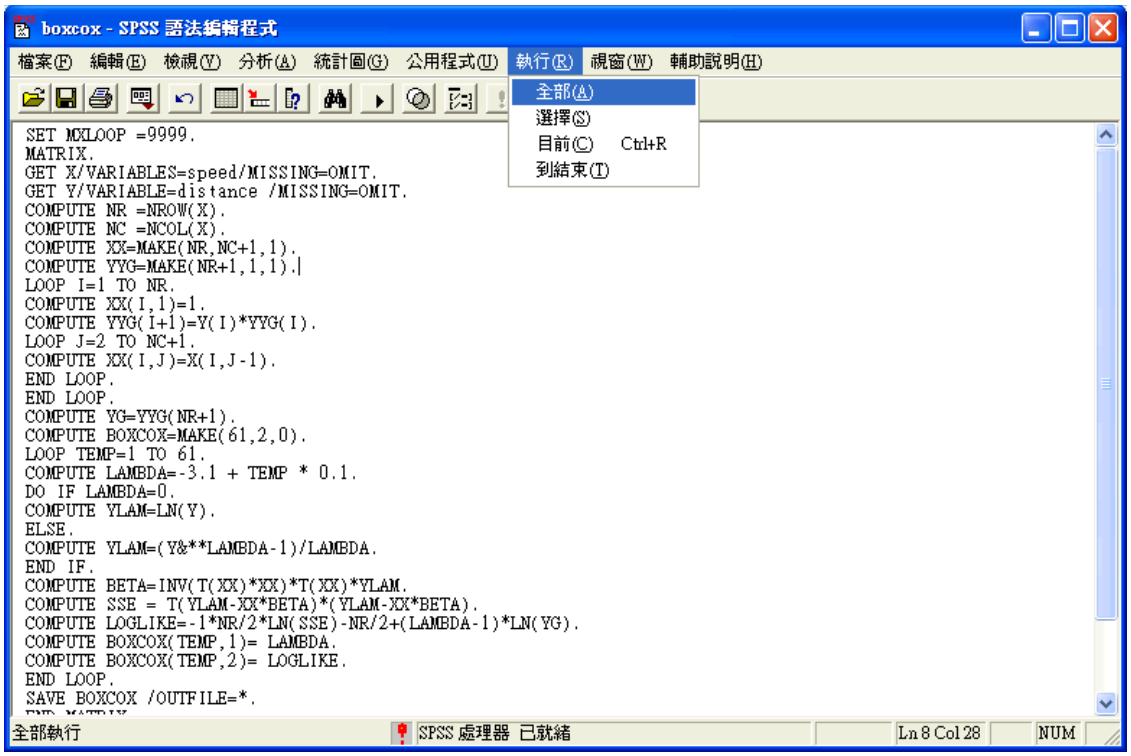
Step 1. SPSS功能列⇒檔案⇒開新檔案⇒語法



Step 2. Box-Cox轉換語法如下，輸入後選擇執行全部即可。

```
SET LENGTH=NONE.
SET MXLOOP = 100000000.
MATRIX.
GET X/VARIABLES=X/MISSING=OMIT.
GET Y/VARIABLE=Y /MISSING=OMIT.
COMPUTE NR =NROW(X).
COMPUTE NC =NCOL(X).
COMPUTE PI=4*ARTAN(1).
COMPUTE XX=MAKE(NR,NC+1,1).
COMPUTE YLAM=MAKE(NR,1,1).
COMPUTE BOXCOX=MAKE(61,2,0).
```

```
COMPUTE YTEMP=0.
LOOP I=1 TO NR.
COMPUTE XX(I,1)=1.
COMPUTE XX(I,2)=X(I).
COMPUTE YTEMP=YTEMP+LN(Y(I)).
END LOOP.
PRINT YTEMP.
LOOP TEMP=1 TO 61.
COMPUTE LAMBDA=-3.1 + TEMP * 0.1.
DO IF LAMBDA=0.
COMPUTE YLAM(:)=LN(Y(:)).
ELSE.
COMPUTE YLAM(:)=(Y(:)&**LAMBDA -1)/LAMBDA.
END IF.
COMPUTE BETA=INV(T(XX)*XX)*T(XX)*YLAM./
COMPUTE MSE = T(YLAM-XX*BETA)*(YLAM-XX*BETA)/NR.
COMPUTE LOGLIKE=-1*NR/2*LN(2*PI)-NR/2*LN(MSE)-1/2+(LAMBDA-1)*(YTEMP).
COMPUTE BOXCOX(TEMP,1)= LAMBDA.
COMPUTE BOXCOX(TEMP,2)= LOGLIKE.
END LOOP.
SAVE BOXCOX /OUTFILE=*.
END MATRIX.
RENAME VARIABLES COL1= LAMBDA COL2= LOGLIKE.
GRAPH
/SCATTERPLOT(BIVAR)=LAMBDA WITH LOGLIKE /MISSING=LISTWISE
/TITLE= 'Log-Likelihood versus lambda'.
```



SPSS報表輸出如下：有下表可發現，當$\lambda = 0.4$時，log-likelihood有最大值。因此最佳之λ值爲0.4，由於不易解釋，因此取$\lambda = 0.5$。

λ	log-likelihood	
0.0	-258.24	
0.1	-254.51	
0.2	-251.86	
0.3	-250.24	
0.4	-249.57	*
0.5	-249.71	
0.6	-250.54	
0.7	-251.92	
0.8	-253.76	
0.9	-255.95	
1.0	-258.43	

經轉換後（$y_{\text{new}} = \sqrt{y}$）之迴歸模式報表輸出如下。由報表可發現模式之解釋能力不僅提高，且由殘差圖亦可發現轉換後資料常態性及恆常性成立。

模式摘要

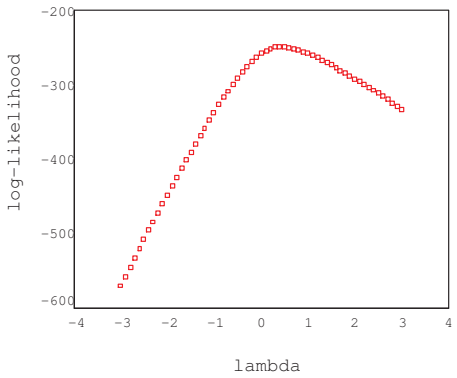
模式	R	R平方	調過後的R平方	估計的標準誤
1	0.842	0.709	0.703	1.1024

變異數分析

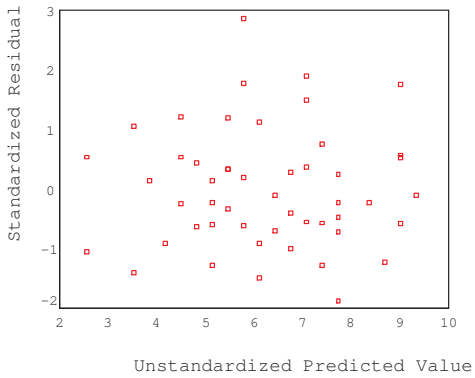
模式		平方和	自由度	平均平方和	F檢定	顯著性
1	迴歸	142.411	1	142.411	117.183	0.000
	殘差	58.334	48	1.215		
	總和	200.745	49			

單一樣本 Kolmogorov-Smirnov 檢定

		Standardized Residual
個數		50
常態參數	平均數	0.0000000
	標準差	0.9897433
最大差異	絕對	0.088
	正的	0.088
	負的	−0.046
Kolmogorov-Smirnov Z 檢定		0.621
漸近顯著性 (雙尾)		0.836



(a)Box-Cox轉換圖



(b)轉換後之殘差圖

5.4 離群值偵測

做資料分析時，常會遇到另一項問題是離群值（outliers）的出現。離群值意指爲資料中一個或多個極大或極小的觀察值。由於迴歸係數容易受到極端值的影響，應該特別注意資料中是否有離群值出現。離群值的出現可能是紀錄者紀錄錯誤，或是資料輸入錯誤。目前離群值的偵測普遍使用Cook於1977年所提出的Cook's Distance來評估觀測值對迴歸參數估計的影響。Cook's Distance運用了第i個觀察值的槓桿作用h_{ii}及殘差e_i來判定觀察值是否具影響力，其定義爲

$$D_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{p \times \text{MSE}} = \frac{e_i^2}{p \times \text{MSE}} \frac{h_{ii}}{(1-h_{ii})^2} = \frac{f_i^2}{p} \frac{h_{ii}}{1-h_{ii}} \tag{5-16}$$

其中

1 $\hat{y}_j$爲第j個預測值；$\hat{y}_{j(i)}$爲移除第i個觀察值所得之第j個預測值

2 h_{ii}爲槓桿值（leverage value），$0 < h_{ii} < 1$，h_{ii}爲矩陣 $\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ 的第i個對角線元素值。當$h_{ii} > 2p/n$時，第i個觀察值可視爲具高影響力。

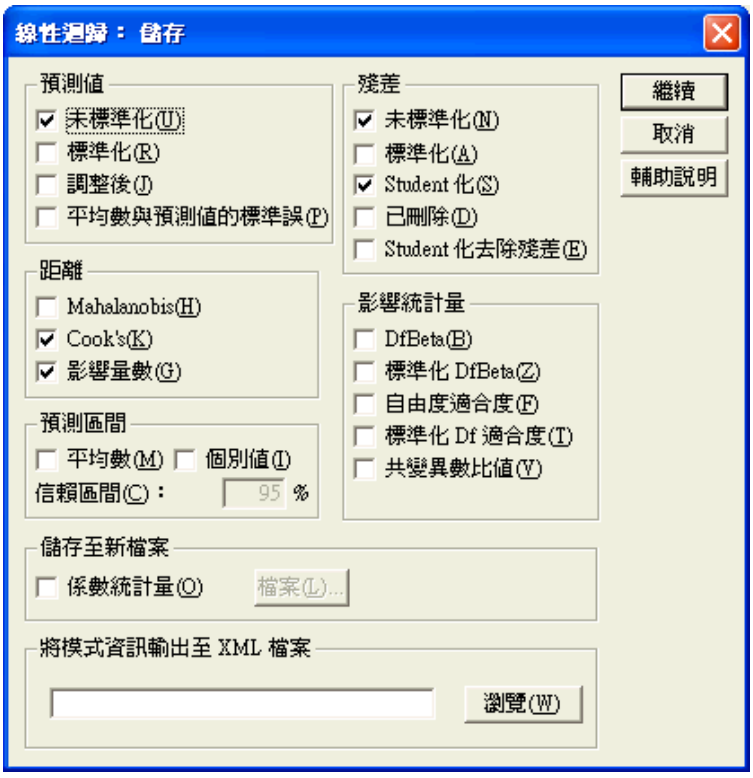
3 $f_i = e_i/[\sqrt{\text{MSE}}(1-h_{ii})]$爲學生化殘差（Studentized residual）

由（5-16）可發現，D_i與h_{ii}呈反相變動；與e_i^2呈正相變動。如果殘差以及（或者）槓桿值很大，則Cook's Distance將很大，我們就可以據此判定該觀察值具影響力。一般來說，當$D_i > 1$時，則可將此觀察值視爲具影響力之離群值，應以予以刪除。

以上述汽車時速與煞車距離資料爲例，轉換後（$y_{\text{new}} = \sqrt{y}$）迴歸模式之槓桿值（$h_{ii}$）及Cook Distance（$D$）如下表，由於$D$值均小於1，因此並無離群值存在。

y_{new}	x	$\hat{y}$	e	h_{ii}	f	D
1.41	4	2.5653	−1.1553	0.1149	−1.1140	0.0537
3.16	4	2.5653	0.5947	0.1149	0.5734	0.0142
2.00	7	3.5328	−1.5328	0.0715	−1.4430	0.0535
4.69	7	3.5328	1.1572	0.0715	1.0895	0.0305
4.00	8	3.8552	0.1448	0.0600	0.1354	0.0004
3.16	9	4.1777	−1.0177	0.0499	−0.9472	0.0157
4.24	10	4.5002	−0.2602	0.0413	−0.2411	0.0008
5.1	10	4.5002	0.5998	0.0413	0.5557	0.0044
5.83	10	4.5002	1.3298	0.0413	1.2320	0.0218
4.12	11	4.8227	−0.7027	0.0341	−0.6486	0.0050
⋮	⋮	⋮	⋮	⋮	⋮	⋮
10.95	24	9.0149	1.9351	0.0740	1.8242	0.0886
9.22	25	9.3374	−0.1174	0.0873	−0.1115	0.0004

若要在SPSS迴歸分析中加入離群值偵測，僅需在迴歸分析對話方塊選取『儲存』，並將『Student化殘差』、『Cook's』及『影響量數』選取即可。操作畫面如下：


線性迴歸：儲存
預測值
未標準化(U)
標準化(R)
調整後(J)
平均數與預測值的標準誤(P)
殘差
未標準化(N)
標準化(A)
Student 化(S)
已刪除(D)
Student 化去除殘差(E)
繼續
取消
輔助說明
距離
Mahalanobis(H)
Cook's(K)
影響量數(G)
影響統計量
DfBeta(B)
標準化 DfBeta(Z)
自由度適合度(F)
標準化 Df 適合度(T)
共變異數比值(V)
預測區間
平均數(M)
個別值(I)
信賴區間(C)：
95
%
儲存至新檔案
係數統計量(O)
檔案(L)...
將模式資訊輸出至 XML 檔案
瀏覽(W)

第 6 章

多元迴歸分析

多元迴歸分析（Multiple Regression Analysis）是簡單相關的一種延伸應用，用以瞭解一個以上解釋變數與應變項的函數關係。而每個預測變項的預測能力，是研究者重要的參考指標。例如教師想了解聯考成績（Y）與各模擬考成績（X_i）之間之關係。當解釋變數僅一個時，稱爲簡單迴歸$Y = \beta_0 + \beta_1 X$；而解釋變數有二個以上時，則稱爲多元迴歸。

6.1 參數估計

當解釋變數有二個以上時，多元迴歸之模式寫成

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

其模式以矩陣表示可寫爲

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I}) \tag{6-1}$$

或可改寫爲

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{6-2}$$

與第五章相同，我們可以普通最小平方法來估計迴歸參數β，即

$$\begin{aligned} \min \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{Y}^\top\mathbf{Y} - \mathbf{Y}^\top\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{Y} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} \end{aligned} \tag{6-3}$$

上式有最小值之必要條件爲

$$\begin{aligned}\frac{\partial \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}}{\partial \boldsymbol{\beta}} &= -\mathbf{Y}^\top \mathbf{X} - \mathbf{X}^\top \mathbf{Y} + 2\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} \\ &= 2\left(\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} - \mathbf{X}^\top \mathbf{Y}\right) \\ &= 0 \end{aligned} \tag{6-4}$$

求解(6.4)可得

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y} \tag{6-5}$$

由於$\boldsymbol{\beta}$是以最小平方法所求得，由高斯馬可夫定理可知，$\boldsymbol{\beta}$亦具備BLUE性質。此外，我們也可推得當$\varepsilon_i \overset{iid}{\sim} \mathbf{N}\left(0, \sigma^2\mathbf{I}\right)$的假設成立時，$\widehat{\beta}$之抽樣分配分別爲

$$\widehat{\boldsymbol{\beta}} \sim \mathbf{N}\left(\boldsymbol{\beta}, \sigma^2\left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\right) \tag{6-6}$$

我們可將總變異以矩陣方式表示爲

$$\begin{aligned}\text{SST} &= \left(\mathbf{Y} - \overline{\mathbf{Y}}\right)^\top \left(\mathbf{Y} - \overline{\mathbf{Y}}\right) \\ &= \mathbf{Y}^\top \mathbf{Y} - \mathbf{Y}^\top \overline{\mathbf{Y}} - \overline{\mathbf{Y}}^\top \mathbf{Y} + \overline{\mathbf{Y}}^\top \overline{\mathbf{Y}} \\ &= \mathbf{Y}^\top \mathbf{Y} - \mathbf{Y}^\top \frac{1}{n}\mathbf{1}\mathbf{Y} - \frac{1}{n}\mathbf{Y}^\top \mathbf{1}^\top \mathbf{Y} + \frac{1}{n}\mathbf{Y}^\top \mathbf{1}^\top \frac{1}{n}\mathbf{1}\mathbf{Y} \\ &= \mathbf{Y}^\top \mathbf{Y} - \frac{2}{n}\mathbf{Y}^\top \mathbf{1}\mathbf{Y} + \frac{1}{n^2}\mathbf{Y}^\top \mathbf{1}^\top \mathbf{1}\mathbf{Y} \\ &= \mathbf{Y}^\top \mathbf{Y} - \frac{1}{n}\mathbf{Y}^\top \mathbf{1}\mathbf{Y} \end{aligned} \tag{6-7}$$

如同前一章節所介紹，我們知道應變數Y之所以有差異，有一部分的原因式來自於解釋變數X所引起，另一部份則爲未知原因所導致，此一部份則是所謂殘差。又眞正的殘差

$$\begin{aligned}\text{SSE} &= \left(\mathbf{Y} - \widehat{\mathbf{Y}}\right)^\top \left(\mathbf{Y} - \widehat{\mathbf{Y}}\right) \\ &= \left(\mathbf{Y} - \mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y}\right)^\top \left(\mathbf{Y} - \mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y}\right) \\ &= \mathbf{Y}^\top \mathbf{Y} - \mathbf{Y}^\top \mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y} - \mathbf{Y}^\top \mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y} \\ &\quad + \mathbf{Y}^\top \mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y} \\ &= \mathbf{Y}^\top \mathbf{Y} - \mathbf{Y}^\top \mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y} \\ &= \mathbf{Y}^\top \mathbf{Y} - \mathbf{Y}^\top \mathbf{X}\boldsymbol{\beta} \end{aligned} \tag{6-8}$$

由於SST = SSR + SSE，故可推得

$$\begin{aligned} \text{SSR} &= \text{SST} - \text{SSE} \\ &= \mathbf{Y}^\top \mathbf{X}\beta - \frac{1}{n}\mathbf{Y}^\top \mathbf{1}\mathbf{Y} \\ &= \mathbf{Y}^\top \mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\mathbf{X}^\top \mathbf{Y} - \frac{1}{n}\mathbf{Y}^\top \mathbf{1}\mathbf{Y} \\ &= \mathbf{Y}^\top \left[\mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\mathbf{X}^\top - \frac{1}{n}\mathbf{1}\right]\mathbf{Y} \end{aligned} \tag{6-9}$$

將各項變異來源編製成下表：

ANOVA Table

變異來源	平方和	自由度	均方和	F值
SSR	$\mathbf{Y}^\top \left[\mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\mathbf{X}^\top - \frac{1}{n}\mathbf{1}\right]\mathbf{Y}$	k	MSR= $\frac{SSR}{k-1}$	F = $\frac{\text{MSR}}{\text{MSE}}$
SSE	$\mathbf{Y}^\top \mathbf{Y} - \mathbf{Y}^\top \mathbf{X}\boldsymbol{\beta}$	$n-1-k$	MSE= $\frac{SSE}{n-1-k}$	
SST	$\mathbf{Y}^\top \mathbf{Y} - \frac{1}{n}\mathbf{Y}^\top \mathbf{1}\mathbf{Y}$	$n-1$		

又母體變異數σ^2未知，且$E(\text{SSE}) = (n-1-k)\sigma^2$，因此MSE在此可作爲母體變異數之不偏估計式。

6.2 β之統計推論

在多元迴歸模式中，爲了解模式是否適用，我們首先必須了解應變數與解釋便是之間是否有線性關係存在？爲求解此一問題，即檢定下列假設，

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1 : \beta_i \text{不全爲} 0 \end{cases}$$

倘若檢定結果不棄卻虛無假設，則表示模式中之解釋變數對應變數並無線性關係存在。此一檢定稱之爲整體模式適合度檢定，其棄卻域爲

$$C = \left\{ F = \frac{\text{MSR}}{\text{MSE}} > F_\alpha(k-1, n-1-k) \right\}$$

當檢定結果爲棄卻虛無假設，則表示β_i不全爲0。因此，研究者則必須進行邊際檢定，以了解各別β_i是否顯著。由於$\widehat{\beta} \sim \mathbf{N}\left(\beta, \sigma^2\left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\right)$，$\text{SSE}/\sigma^2 \sim \chi^2(n-1-k)$及$\sigma^2$未知，由$t$分配之定義可知

$$\frac{\frac{\widehat{\beta}_i - \beta_i}{\sqrt{\sigma^2 C_{ii}}}}{\sqrt{\frac{\text{SSE}/\sigma^2}{n-1-k}}} = \frac{\widehat{\beta}_i - \beta_i}{\sqrt{\text{MSE} \times C_{ii}}} \sim t\,(n-1-k) \tag{6-10}$$

其中C_{ii}爲$\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$第$i$個對角線元素值。故由可求得$\beta_i$之$(1-\alpha)\times 100\%$信賴區間爲

$$\left[\widehat{\beta}_i - t_{\frac{\alpha}{2}}\,(n-1-k)\,\sqrt{\mathrm{MSE}\times C_{ii}}, \widehat{\beta}_i + t_{\frac{\alpha}{2}}\,(n-k)\,\sqrt{\mathrm{MSE}\times C_{ii}}\right]$$

當統計假設爲雙尾檢定時，即

$$\begin{cases} H_0 : \beta_i = k \\ H_1 : \beta_i \neq k \end{cases}$$

其棄卻域爲

$$C = \left\{ \mid \widehat{\beta}_i - k \mid > t_{\frac{\alpha}{2}}\,(n-1-k)\,\sqrt{\mathrm{MSE}\times C_{ii}} \right\}$$

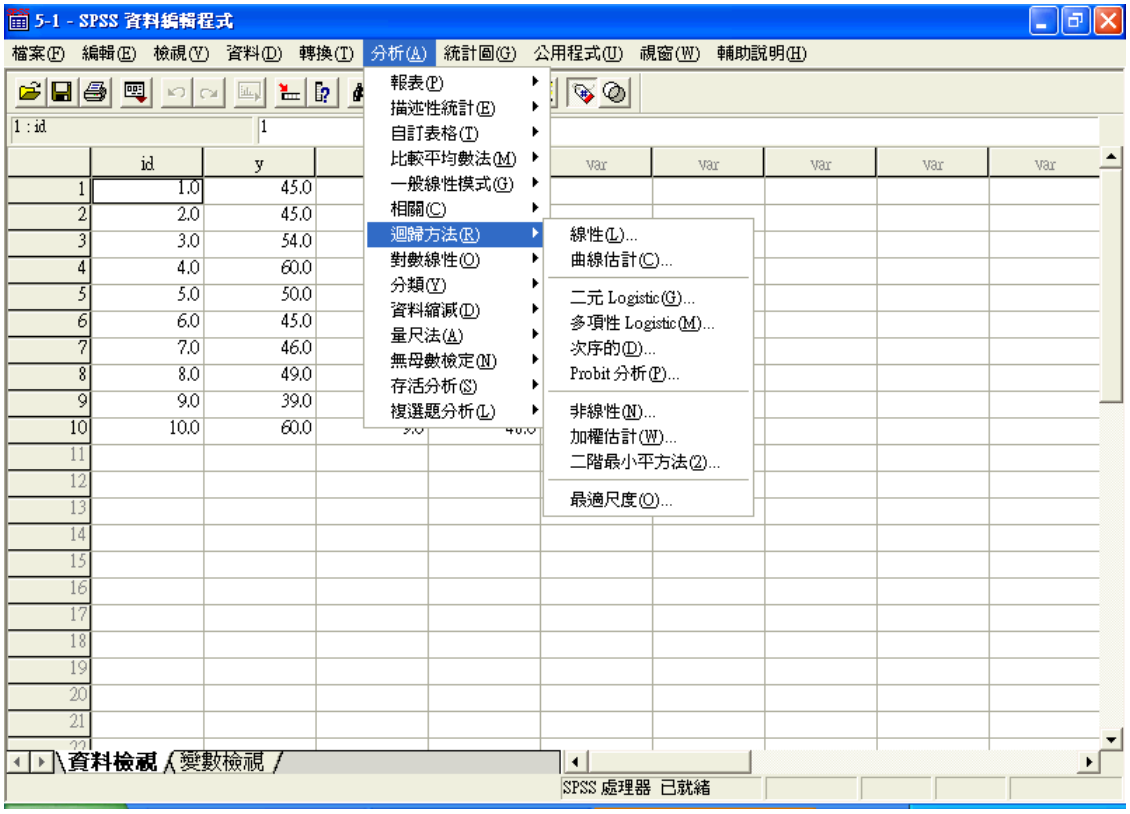
例 6-1. 爲了解Y與X_1及X_2之關係，乃從事一項研究，所得樣本如下：

Y	45	45	54	60	50	45	46	49	39	60
X_1	7	6	9	8	9	12	12	11	13	9
X_2	62	62	45	40	55	58	70	64	63	48

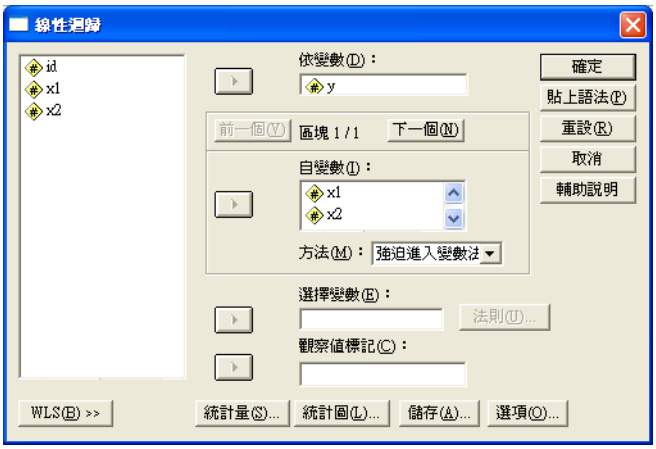
試求迴歸估計式。

解：以SPSS爲求解工具，操作步驟如下：

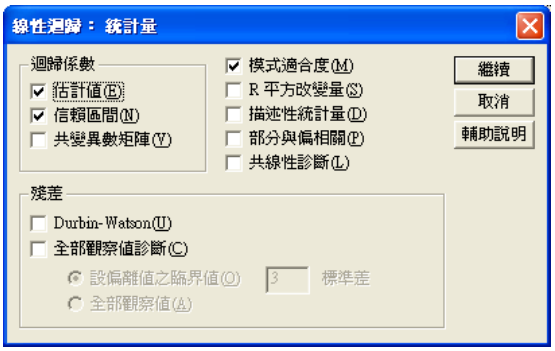
Step 1. SPSS功能列⇒分析⇒迴歸方法⇒線性



Step 2. 選擇依變數與自變數



Step 3. 選取『統計量』對話方塊，並勾選『信賴區間』



SPSS報表輸出如下：

模式摘要

模式	R	R平方	調過後的R平方	估計的標準誤
1	0.833	0.694	0.607	4.303

首先，由『模式摘要』表可知$R^2 = 0.694$，即Y之變異有69.4%可由X_1及X_2所解釋。上表中有一欄爲『調過後的R平方』，其目的在修正迴歸模型中不斷增加獨變項，R^2不會減低之情況。研究者爲了提高模型的解釋力，不斷的投入解釋變數；每增加一個解釋變數，損失一個自由度。最後模型中無關的解釋變數過多，反而失去了計算的效率。因此，爲了懲罰增加解釋變數所損失的計算效率，將自由度的變化作爲分子與分母項的除項加以控制，可以反應因爲獨變項數目變動的簡效性損失的影響。調整後判定係數公式如下：

$$\text{Adj } R^2 = R^2 - \frac{k-1}{n-k}(1-R^2) \tag{6-11}$$

所以，調整的R^2乃模型配適能力與模型複雜度二者折衷後的結果，其數值不一定會隨著解釋變數數目的增加而遞增。

變異數分析

模式		平方和	自由度	平均平方和	F檢定	顯著性
1	迴歸	294.477	2	147.238	7.951	0.016
	殘差	129.623	7	18.518		
	總和	424.100	9			

由於『變異數分析』報表中P-value $= 0.0158 < 0.05$，因此棄卻虛無假設

$$\begin{cases} H_0 : \beta_1 = \beta_2 = 0 \\ H_1 : \beta_1, \beta_2 \text{不全爲} 0 \end{cases}$$

即β_i不全爲0。又由下表可發現，Y與X_1和X_2之迴歸估計式爲

$$Y = 84.1605 - 0.2347X_1 - 0.5751X_2$$

此外， X_1係數之檢定未達顯著水準，因此X_1對Y並無貢獻。反之，由於X_2係數之檢定達顯著水準，因此，X_2對Y具貢獻性。

係數

模式		未標準化係數 β之估計值	標準誤	t	顯著性	95% 信賴區間	
						下限	上限
1	(常數)	84.160	9.126	9.222	0.000	62.581	105.740
	X1	−0.235	0.666	−0.352	0.735	−1.809	1.340
	X2	−0.575	0.162	−3.560	0.009	−0.957	−0.193

6.3 共線性診斷

在決定迴歸分析的模式後，我們應進一步檢視自變項間是否有多元共線性（Multicollinearity）的問題，也就是自變項間是否有高度相關的問題。如果自變項間高度相關的話，不但迴歸係數會變得不穩定外，迴歸方程式的解釋變數也會相互削弱各自對應變數的邊際影響。這會使得本身的迴歸係數變小，於是會出現整體的迴歸方程式顯著，但個別解變數的迴歸係數（β_i值很小）卻不顯著的現象，即在總檢定中

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1 : \beta_i \text{不全爲} 0 \end{cases}$$

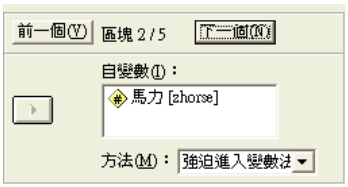
棄卻虛無假設，但在邊際檢定中

$$\begin{cases} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0 \end{cases}$$

不棄卻虛無假設。此種情形使得迴歸分析結果難以解釋，稱爲共線性。

共線性之所以發生，主要是研究者爲了提高模型的解釋力，不斷的投入解釋變數。當所投入之解釋變數之間具有高度相關時，則矩陣$\mathbf{X}^\top\mathbf{X}$中的某兩行或某兩列就會近似。因此，$(\mathbf{X}^\top\mathbf{X})^{-1}$中元素就會變大。由變異數來源表中，讀者即可發現此時$\mathbf{Y}^\top\left[\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top - \frac{1}{n}\mathbf{1}\right]\mathbf{Y}$因$(\mathbf{X}^\top\mathbf{X})^{-1}$中元素變大導致SSR變大；同時SSE則變小。此種結果便導致F因太大落入棄卻域而棄卻虛無假設。又由（6-10）可知，當$(\mathbf{X}^\top\mathbf{X})^{-1}$中元素變大也導致邊際檢定之$t$值變小，因而未落入棄卻域。

以SPSS內部檔案cars.sav爲例，若要分析排氣量、馬力、重量、加速時間及氣缸數對行駛里程之影響，讀者可重複上述迴歸分析操作步驟。將行駛里程輸入依變數後，首先將排氣量輸入至自變數。接著選擇『下一個』，並依序將馬力、重量、加速時間及氣缸數輸入至自變數。



SPSS就會依序將排氣量、馬力、重量、加速時間及氣缸數加入至自變數中進行迴歸分析。由『模式摘要表』中可發現，每加入一個新變數後， 判定係數R^2就會增加。

模式摘要

模式	R	R^2	調過後的 R 平方	估計的標準誤
1	0.805^a	0.648	0.647	0.5942720
2	0.814^b	0.662	0.661	0.5825584
3	0.839^c	0.704	0.702	0.5458104
4	0.839^d	0.704	0.701	0.5465001
5	0.840^e	0.705	0.701	0.5465573

^{a} 預測變數：(常數), 排氣量
^{b} 預測變數：(常數), 排氣量, 馬力
^{c} 預測變數：(常數), 排氣量, 馬力, 重量
^{d} 預測變數：(常數), 排氣量, 馬力, 重量, 加速時間
^{e} 預測變數：(常數), 排氣量, 馬力, 重量, 加速時間, 氣缸數

然而由最後模式之係數檢定表中可以發現，大多數的解釋變數都未達顯著水準。

係數

模式		未標準化係數 β之估計值	標準誤	t	顯著性	95% 信賴區間 下限	95% 信賴區間 上限
5	(常數)	0.000	0.028	0.000	1.000	−0.054	0.054
	排氣量	−0.002	0.124	−0.019	0.985	−0.246	0.241
	馬力	−0.222	0.084	−2.628	0.009	−0.387	−0.056
	重量	−0.564	0.089	−6.341	0.000	−0.739	−0.389
	加速時間	−0.010	0.045	−0.216	0.829	−0.099	0.080
	氣缸數	−0.087	0.090	−0.959	0.338	−0.264	0.091

共線性沒有存不存在的問題，而是程度的問題，應盡量消除它。消除共線性的方法主要有目視法、VIF法、逐步迴歸法及脊迴歸法（Ridge Regression）。分別介紹如下：

1 目視法：將解釋變數間具有高度相關者刪除，通常以相關係數絕對值大於0.7即爲高相關。主要原因爲若兩變數呈高相關，則表示其中一變數可被另一變數解釋比例亦會較高。因此，可從此二變數中剔除一變數以避免共線性發生。

2 VIF法：變異數膨脹因素（Variance Inflation Factor；VIF）爲容忍度的倒數，VIF的值愈大，表示自變項的容忍度愈小，愈可能發生共線性。容忍度定義爲$1-R_i^2$，其中R_i^2爲模式

$$X_i = \gamma_0 + \gamma_1 X_1 + \cdots + \gamma_{i-1} X_{i-1} + \gamma_{i+1} X_{i+1} + \cdots + \gamma_k X_k$$

之判定係數。同理，若R_i^2高，則表示變數X_i可以被其他變數所解釋的比例也會很高。因此，可以從模式中剔除，以避免共線性發生。爲避免各解釋變數間的共線性造成迴歸分析結果的偏誤，通常若VIF大於10則表示該解釋變數與其他解釋變數存在共線性的問題。SPSS之共線性診斷所提供報表尙有提供解釋變數間相關係數矩陣之特徵值及條件指標。若相關係數矩陣之特徵值大，則表示解釋變數存在著較強的線性關係，可藉由變數剔除已達構面縮減之目的，原理我們將在下一章主成份分析中介紹。其中條件指標之定義如下：

$$\text{條件指標}_i = \sqrt{\frac{\text{最大特徵值}}{\text{第}i\text{大的特徵值}}} \tag{6-12}$$

3 向前選取法（Forward）：預測變項的取用順序，以具有最大預測力且達統計顯著水準的獨變項首先被選用，然後依序納入方程式中，直到所有達顯著的預測變項均被納入迴歸方程式。

4 向後選取法（Backword）：與順向進入法相反的程序，所有的預測變項先以同時分析法的方式納入迴歸方程式的運算當中，然後逐步的將未達統計顯著水準的預測變項，以最弱、次弱的順序自方程式中予以排除。直到所有未達顯著的預測變項均被淘汰完畢爲止。

5 逐步迴歸法（Stepwise）：首先模式中不包含任何預測變項。然後採順向選擇法，根據對模式的貢獻最大者，挑選預測變項進入迴歸模式中。而在每一步驟中，已被納入模式的預測變項則必須再經過反向淘汰法的考驗，以決定該變項要被淘汰亦或留下。所有的預測變項並非同時被取用來進行預測，而是依據解釋力的大小，逐步的檢視每一個預測變項的影響。

6 脊迴歸法：其構想是對標準化後的解釋變數矩陣$\mathbf{X}$，將$\mathbf{X}^\top\mathbf{X}$矩陣的對角線元素上加上一個微小的正數$k$。藉由適當$k$值的選擇以來穩定迴歸係數$\beta$的估計量，降低不當資訊對係數估計的影響。令$\beta_{RR}$爲脊迴歸所求得之估計式，則

$$\widehat{\boldsymbol{\beta}}_{\text{RR}} = (\mathbf{X}^\top\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{Y} \tag{6-13}$$

在此必須注意的是，脊迴歸法所求得之估計式，$\widehat{\boldsymbol{\beta}}_{\text{RR}}$，並非$\boldsymbol{\beta}$之不偏之估計式。脊迴歸法雖可避開變異數方面的缺失，卻必須以『偏差』（bias）爲代價；即k增加，則R^2下降及偏差及增加。因此，$\boldsymbol{\beta}$之估計過程中犧牲一點不偏性，但獲得有效性（較小的估計標準誤，$\text{Var}(\beta_{\text{RR}}) < \text{Var}(\beta)$）。

例 6-2. 以SPSS內部檔案cars.sav爲例，試分析排氣量、馬力、重量、加速時間及氣缸數對行駛里程之影響，並建立適當之迴歸模式，$\alpha = 0.05$。

解：由於各解釋變數單位不一，爲消除單位影響，故將各變數予以標準化。標準化後各變數之相關係數矩陣如下，由表可發現變數之間具高度相關，因此可能存在共線性。

相關

		行駛里程	排氣量	馬力	重量	加速時間	氣缸數
行駛里程	Pearson 相關	1.000	−0.805	−0.776	−0.831	0.431	−0.776
	顯著性 (雙尾)	.	0.000	0.000	0.000	0.000	0.000
排氣量	Pearson 相關	−0.805	1.000	0.898	0.934	−0.548	0.951
	顯著性 (雙尾)	0.000	.	0.000	0.000	0.000	0.000
馬力	Pearson 相關	−0.776	0.898	1.000	0.863	−0.701	0.842
	顯著性 (雙尾)	0.000	0.000	.	0.000	0.000	0.000
重量	Pearson 相關	−0.831	0.934	0.863	1.000	−0.425	0.897
	顯著性 (雙尾)	0.000	0.000	0.000	.	0.000	0.000
加速時間	Pearson 相關	0.431	−0.548	−0.701	−0.425	1.000	−0.511
	顯著性 (雙尾)	0.000	0.000	0.000	0.000	.	0.000
氣缸數	Pearson 相關	−0.776	0.951	0.842	0.897	−0.511	1.000
	顯著性 (雙尾)	0.000	0.000	0.000	0.000	0.000	.

由於各變數皆經標準化轉換，故以下迴歸模式之常數項可忽略。

SPSS之操作步驟如下：

Step 1. SPSS功能列⇒分析⇒描述性統計⇒描述性統計量，藉由描述性統計量將解釋變數標準化


car - SPSS 資料編輯程式
檔案(F) 編輯(E) 檢視(V) 資料(D) 轉換(T) 分析(A) 統計圖(G) 公用程式(U) 視窗(W) 輔助說明(H)
報表(P)
描述性統計(E)
自訂表格(T)
比較平均數法(M)
一般線性模式(G)
相關(C)
迴歸方法(R)
對數線性(O)
分類(Y)
資料縮減(D)
量尺法(A)
無母數檢定(N)
存活分析(S)
複選題分析(L)
次數分配表(F)...
描述性統計量(D)...
預檢資料(E)...
交叉表(C)...
22 : zcylinde
-0.8604474805
zmpg zengine y1 y2 Var
2 -.70468 -1.18638 795.64 1876.70
3 -.57616 -1.18638 867.97 2057.29
4 -.25485 -1.09081 1009.60 2401.05
5 1.09466 -.56515 880.98 2040.07
6 .77335 -.92833 699.34 1620.78
7 .32351 -.69895 867.97 1980.20
8 -.25485 -.69895 989.30 2303.10
9 .83761 -.56515 911.74 2119.65
10 1.60876 -.56515 906.76 2106.46
11 1.37742 -.85187 828.81 1952.99
12 1.28745 -.41223 -.37209 -.49312 -.84608 -.86045 979.72 2272.56
13 .19499 -.69895 .28102 -.35716 -.73732 -.86045 1015.78 2365.17
14 .06646 -.57470 -.21534 -.32051 -.73732 -.86045 1023.42 2397.33
15 .38778 -.51736 -.84233 -.48248 -.70106 -.86045 969.09 2280.87
16 2.71407 -.98567 -.97296 -1.32781 -.62856 -.86045 703.67 1642.58
17 1.09466 -.47913 -.21534 -.36426 -.59230 -.86045 1016.04 2366.68
18 1.08181 -1.00479 -.86846 -1.23914 -.55605 -.86045 729.82 1707.81
19 .70909 -.99523 -.89458 -1.22495 -.55605 -.86045 733.85 1718.70
20 .58056 -.74674 -.37209 -1.00505 -.55605 -.86045 818.17 1884.79
21 -.06206 -.68939 -.47659 -.89037 -.55605 -.86045 851.66 1971.88
22 .19499 -.77541 -.24147 -.88091 -.55605 -.86045 854.27 1976.10
資料檢視 變數檢視
描述性統計量
SPSS 處理器 已就緒

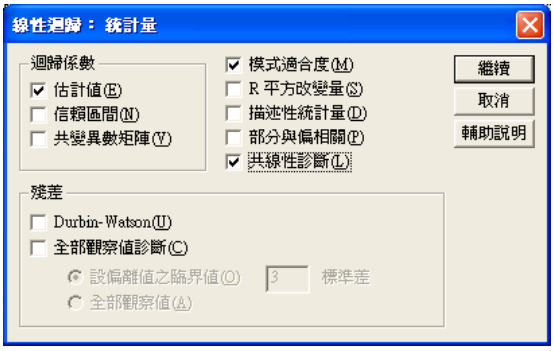
Step 2.


描述性統計量
馬力 [horse]
重量 [weight]
加速時間 [accel]
型式年份 (modulo
製造地區 [origin]
氣缸數 [cylinder]
變數(V):
行駛里程 [zmpg]
排氣量 [zengine]
馬力 [zhorse]
重量 [zweight]
加速時間 [zaccel]
氣缸數 [zcylinde]
確定
貼上語法(P)
重設(R)
取消
輔助說明
將標準化的數值存成變數(Z)
選項(O)...

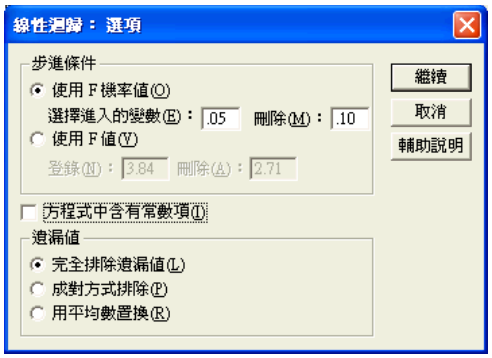
Step 3. 以標準化之解釋變數建立迴歸模式



Step 4. 在『統計量』方塊中，選取輸出『共線性診斷表』



Step 5. 在『選項』方塊中， 將『方程式中含有常數項』取消



SPSS輸出之報表如下：

模式摘要

模式	R	R平方	調過後的R平方	估計的標準誤
1	0.840	0.705	0.701	0.5465573

變異數分析

模式		平方和	自由度	平均平方和	F檢定	顯著性
1	迴歸	274.991	5	54.998	184.588	0.000
	殘差	115.009	386	0.298		
	總和	390.000	391			

係數

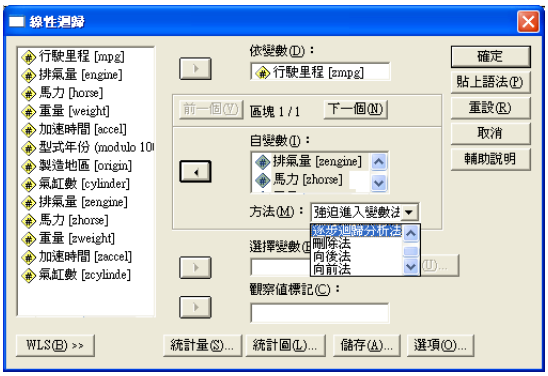
		未標準化係數				共線性統計量	
模式		β之估計值	標準誤	t	顯著性	允差	VIF
1	排氣量	−0.0023	0.124	−0.019	0.985	0.050	20.019
	馬力	−0.2215	0.084	−2.632	0.009	0.108	9.273
	重量	−0.5638	0.089	−6.350	0.000	0.097	10.319
	加速時間	−0.0098	0.045	−0.216	0.829	0.371	2.698
	氣缸數	−0.0866	0.090	−0.960	0.338	0.094	10.658

由『係數』表中發現，排氣量、重量氣及缸數由於VIF值過高，因此容易在成共線性，在此可將此三解釋變數予以剔除。而由以下『共線性診斷』表發現特徵值大，則表示解釋變數存在著較強的線性關係，可藉由剔除變異數比例較大之變數，如排氣量及氣缸數，已達構面縮減。

共線性診斷

				變異數比例				
模式	維度	特徵值	條件指標	排氣量	馬力	重量	加速時間	氣缸數
1	1	4.0820	1.000	0.00	0.01	0.01	0.01	0.01
	2	0.6870	2.437	0.00	0.00	0.02	0.42	0.01
	3	0.1310	5.573	0.01	0.30	0.09	0.07	0.32
	4	0.0639	7.993	0.00	0.62	0.74	0.50	0.08
	5	0.0357	10.691	0.98	0.07	0.14	0.00	0.58

最後我們以逐步迴歸法選取解釋變數並進行迴歸分析。讀者可在迴歸分析之解釋變數選取方法中選擇以『逐步迴歸分析法』進行。需注意的是，在逐步迴歸法中，並不需要進行共線性診斷。其原因爲解釋變數在篩選過程中，解釋能力較低者（即容易被其他變數取代者）將被排除在模式之外。SPSS選擇視窗如下：



由輸出之報表可知『重量』爲最具解釋力之變數，因此最快進入模式。讀者亦可由相關矩陣中發現，『重量』與行駛里程之相關性在所有解釋變數中最高。在第二個模式中，進入變數爲馬力。讀者可由以下模式摘要表中發現，由第一個模式再新增一解釋變數所增加之解釋能力僅提高1.4%。因此，即使在新增其他解釋變數，對整體模式之解釋能力的幫助有限。爲不降低模式之效率，其他之解釋變數故不加入模式中。

模式摘要

模式	R	R平方	調過後的R平方	估計的標準誤
1	0.831	0.690	0.689	0.5568194
2	0.839	0.704	0.702	0.5449659

[1]解釋變數：重量
[2]解釋變數：重量, 馬力

變異數分析

模式		平方和	自由度	平均平方和	F檢定	顯著性
1	迴歸	269.081	1	269.081	867.870	0.000
	殘差	120.919	390	0.310		
	總和	390.000	391			
2	迴歸	274.472	2	137.236	462.093	0.000
	殘差	115.528	389	0.297		
	總和	390.000	391			

[1]解釋變數：重量
[2]解釋變數：重量, 馬力

係數

		未標準化係數			
模式		β之估計值	標準誤	t	顯著性
1	重量	−0.831	0.028	−29.460	0.000
2	重量	−0.630	0.055	−11.534	0.000
	馬力	−0.233	0.055	−4.260	0.000

在變異數分析表中，讀者可發現，總和之自由度=樣本數。其原因爲變數皆經由標準化轉換，因此$\mu_{\text{行駛里程}} = 0$並不需另外估計，因此自由度不需減一之故。最後，由模式二之係數表可知，最適迴歸模式爲

$$\text{行駛里程} = -0.630 \times \text{重量} - 0.233 \times \text{馬力}$$

此外，由於SPSS並無脊迴歸之操作介面，若要執行脊迴歸分析，則必須藉由執行程式語法之功能才可執行。SPSS操作步驟如下：

Step 1. SPSS功能列⇒檔案⇒開新檔案⇒語法

Step 2. 脊迴歸執行語法如下，輸入後選擇執行全部即可。

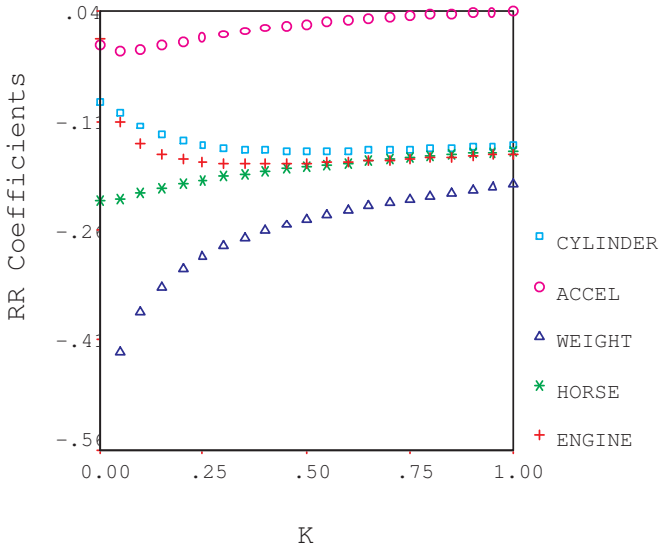
```
INCLUDE 'C:\Program Files\SPSS\Ridge regression.sps'.
RIDGEREG DEP=mpg /ENTER = engine horse weight accel cylinder.
```

SPSS報表輸出如下，由脊迴歸圖可看出，當$k > 0.6$時，各估計值有穩定跡象，因此此例可取$k = 0.6$。若要列出得$k = 0.6$時之迴歸模式報表，可重複步驟2，並以下列語法執行。

```
INCLUDE 'C:\Program Files\SPSS\Ridge regression.sps'.
RIDGEREG DEP=mpg /ENTER = engine horse weight accel cylinder/k=0.6.
```

R^2 AND β COEFFICIENTS FOR ESTIMATED VALUES OF K

K	RSQ	ENGINE	HORSE	WEIGHT	ACCEL	CYLINDER
0.00	0.7051	−0.0023	−0.2215	−0.5638	−0.0098	−0.0866
0.05	0.7029	−0.1144	−0.2194	−0.4301	−0.0196	−0.1017
0.10	0.7003	−0.1453	−0.2121	−0.3731	−0.0164	−0.1209
0.15	0.6979	−0.1590	−0.2051	−0.3391	−0.0110	−0.1335
0.20	0.6956	−0.1661	−0.1991	−0.3157	−0.0054	−0.1417
0.25	0.6933	−0.1698	−0.1940	−0.2982	−0.0002	−0.1471
0.30	0.6911	−0.1718	−0.1896	−0.2843	0.0045	−0.1506
0.35	0.6889	−0.1726	−0.1858	−0.2728	0.0088	−0.1529
0.40	0.6867	−0.1727	−0.1823	−0.2631	0.0126	−0.1543
0.45	0.6844	−0.1723	−0.1792	−0.2546	0.0160	−0.1550
0.50	0.6821	−0.1716	−0.1763	−0.2472	0.0191	−0.1553
0.55	0.6798	−0.1707	−0.1737	−0.2405	0.0219	−0.1553
0.60	0.6775	−0.1696	−0.1712	−0.2345	0.0244	−0.1550
0.65	0.6751	−0.1684	−0.1688	−0.2290	0.0266	−0.1544
0.70	0.6727	−0.1671	−0.1666	−0.2239	0.0286	−0.1538
0.75	0.6703	−0.1658	−0.1645	−0.2192	0.0304	−0.1530
0.80	0.6678	−0.1644	−0.1624	−0.2149	0.0321	−0.1521
0.85	0.6653	−0.1630	−0.1605	−0.2108	0.0335	−0.1512
0.90	0.6628	−0.1616	−0.1586	−0.2070	0.0349	−0.1502
0.95	0.6603	−0.1602	−0.1568	−0.2033	0.0361	−0.1492
1.00	0.6578	−0.1588	−0.1551	−0.1999	0.0372	−0.1481



模式摘要

模式	R	R平方	調過後的R平方	估計的標準誤
$k = 0.6$	0.8231	0.6775	0.6733	4.4473

ANOVA table

	df	SS	MS
Regress	5	15995.22	3199.05
Residual	385	7614.55	19.78

Variables in the Equation

	B	SE(B)	Beta	B/SE(B)
ENGINE	−0.0126	0.0008	−0.1696	−16.4403
HORSE	−0.0348	0.0026	−0.1712	−13.4027
WEIGHT	−0.0022	0.0001	−0.2345	−18.3119
ACCEL	0.0687	0.0491	0.0244	1.3998
CYLINDER	−0.7079	0.0593	−0.1550	−11.9464
Constant	38.7722	1.1012	0.0000	35.2078

第 7 章

主成份分析

在多元迴歸模式中，若解釋變數增加則對模式中的被解釋變數的解釋能力也會增加。然而，如同我們在前一章節所言，解釋變數的數目過多或是彼此間存在高度相關性，容易使得模式對於應變數所提供的訊息有所重疊，嚴重則可能發生共線性，影響模式估計。 是否有一種統計方法能夠將解釋變數的數目減少，且而不降低模式對應變數的解釋能力，如果有，則可使模式更穩定且更具有統計意義。在此，我們將介紹「主成份分析」（Principal Components Analysis ），它的目的即是希望用較少的變數去解釋原始資料的大部分變異。此外，它所提供的新變數也互相獨立，因而可避免後續迴歸分析中所可能發生之共線性。

「主成份分析」首先是由Pearson於1901年所提出的，再由Hotelling（1933）加以發展的一種統計方法。其主要目的是在於將許多變數予以減少，並使其改變爲少數幾個互相獨立的線性組合變數。另外爲了讓主成份盡量保留原始變數所含有的訊息，一般希望經由線性組合而得的主成份之變異數越大越好，使得受試者之間的差異或變化能在這些主成份上保留下來。

7.1 主成份模式

主成份分析在於「轉換」原始變數使其成爲互相獨立的線性組合變數；經由線性組合所求得的主成分仍保有原變數最多的資訊。在此，假設$\mathbf{X}$爲一$p \times 1$的隨機向量，$\mathbf{X}^\top = [X_1, X_2, \dots X_p]^\top$，其中$\mathrm{E}\mathbf{X} = \mu$且$\mathrm{Cov}(\mathbf{X}) = \mathbf{\Sigma}$。由於$\mathbf{\Sigma}$爲一正定矩陣有$p$個特徵值，依序爲$\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$。令

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{bmatrix} = \begin{bmatrix} a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ \vdots \\ a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p \end{bmatrix} \tag{7-1}$$

由於主成份分析希望經由線性組合而得的主成份之變異數越大越好，我們的目標即爲

$$\max \quad Var(Y_1) = \mathbf{a}_1^\top \mathbf{\Sigma} \mathbf{a}_1 \quad \text{subject to} \quad \mathbf{a}_1^\top \mathbf{a_1} = 1 \tag{7-2}$$

在此，我們可以Lagrange乘數法求解上述最大化問題，令

$$L = \mathbf{a}_1^\top \mathbf{\Sigma} \mathbf{a_1} - \lambda \mathbf{a}_1^\top \mathbf{a}_1 \tag{7-3}$$

上式最大化之必要條件爲

$$\frac{\partial L}{\partial \mathbf{a}_1} = 2\Sigma \mathbf{a}_1 - 2\lambda \mathbf{a}_1 = 0 \Leftrightarrow \mathbf{\Sigma} \mathbf{a}_1 = \lambda \mathbf{a}_1 \tag{7-4}$$

可推得$\mathbf{a}_1$必須爲$\mathbf{\Sigma}$之特徵向量。所以，第一個主成份Y_1爲$\mathbf{X}$與最大特徵值λ_1所對應特徵向量$\mathbf{a}_1$之線性組合，我們稱此一線性組合所得之分數爲主成份分分數；Y_2爲$\mathbf{X}$與第二大特徵值λ_2所對應特徵向量$\mathbf{a}_2$之線性組合，以此類推。由於$\mathbf{a} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p]$爲特徵向量所構成之矩陣，對所有$i \neq j$時，$\mathbf{a}_i^\top \mathbf{a}_j = 0$。所以我們可發現主成份分析除了具有幾何學上正交的特性（geometrically orthogonal）外，亦具有統計上正交（statistically orthogonal）的特性，亦即各主成份間彼此無關；這正是主成份分析能解決共線性問題之原因。

由於個別主成份Y_i爲$\mathbf{X}$的線性組合，所以可求得

$$\text{Var}\,(Y_i) = \mathbf{a}_i^\top \mathbf{\Sigma} \mathbf{a}_i = \mathbf{a}_i^\top \mathbf{a} \mathbf{\Lambda} \mathbf{a}^\top \mathbf{a}_i = \lambda_i \tag{7-5}$$

及

$$\begin{aligned} \text{Cov}\,(Y_i, X_j) &= \mathbf{a}_i^\top \mathbf{\Sigma} \mathbf{d}_j = \mathbf{a}_i^\top \mathbf{a} \mathbf{\Lambda} \mathbf{a}^\top \mathbf{d}_j \\ &= a_{ij} \lambda_i \end{aligned} \tag{7-6}$$

其中

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix} \text{及} \mathbf{d}_j = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \text{第}j\text{行}$$

可發現第i個主成份的變異數即爲第i大的特徵值。此外，第i個主成份與第j個變數之相關係數爲

$$\rho_{y_i x_j} = \frac{a_{ij}\lambda_i}{\sqrt{\lambda_i}\sigma_j} = \frac{a_{ij}\sqrt{\lambda_i}}{\sigma_j} \tag{7-7}$$

我們稱此一相關係數爲主成份負荷（loading）。第i個主成份解釋變數X_j, $j = 1, 2, \ldots, j$的變異數比例稱爲**共同性**（Communality），以符號h_i^2表示。由判定係數之定義可知

$$h_i^2 = \sum_{i}^{j} \frac{a_{ij}^2 \lambda_i}{\sigma_j^2} \tag{7-8}$$

又由總變異估計量trace($\mathbf{\Sigma}$)可知

$$\text{trace}\left(\mathbf{\Sigma}\right) = \text{tr}\left(\mathbf{a}\mathbf{\Lambda}\mathbf{a}^\top\right) = \text{trace}\left(\mathbf{a}^\top\mathbf{a}\mathbf{\Lambda}\right) = \text{tr}\left(\mathbf{\Lambda}\right) = \sum_{i=1}^{p}\lambda_i \tag{7-9}$$

所有特徵值之和爲變異數矩陣之跡(trace)，因此，第i個主成份解釋的變異數比例爲

$$\frac{\lambda_i}{\sum_{i=1}^{p}\lambda_i} \tag{7-10}$$

若以相關係數矩陣作爲分解對象時，$\rho_{y_i x_j} = a_{ij}\sqrt{\lambda_i}$；特徵值之和除了爲相關係數矩陣之跡（trace）外，亦爲變項之個數。

7.2 主成份的個數選取

主成份分析的目的是希望用較少的變數去解釋原始資料的大部分變異。當然，若將所有的主成份皆選取，則無法達到資料構面縮減的目的。然而，到底需要幾個主成份才足夠，這沒有一定的標準答案。一般說來，主成份的個數選取，有以下幾種方法：

1 取特徵值大於全部平均值者。因爲特徵值大於平均數，表示在解釋總變異的貢獻優於一般。

2 若資料已經標準化，由於總變異爲變數個數，故取特徵值大於1者。

3 將特徵值由大至小排列畫成陡坡圖，並選取開始變平緩的點所對應的個數

在此必須注意的是，若原有變數之間的相關性很低，那麼就不需要作主成份分析。原因是經過主成份分析後所選取的主成份個數基本上不會與原有變數的個數相差多少。Bartlett's test of sphericity可以檢定相關係數矩陣是否爲單位矩陣，其檢定統計量爲

$$\begin{cases} H_0 : \rho = \mathbf{I} \\ H_1 : \rho \neq \mathbf{I} \end{cases}$$

及

$$\chi^2 = -\left[n - 1 - \frac{2p+5}{6}\right] \times \ln(|R|) \sim \chi^2\left(\frac{p(p-1)}{2}\right) \tag{7-11}$$

7.3 範例

以SPSS內部檔案cars.sav爲例，試以排氣量、馬力、重量、加速時間及氣缸數進行主成份分析。

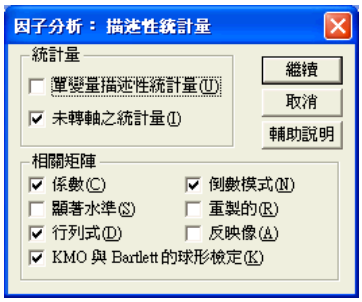
Step 1. SPSS功能列⇒分析⇒資料縮減⇒因子



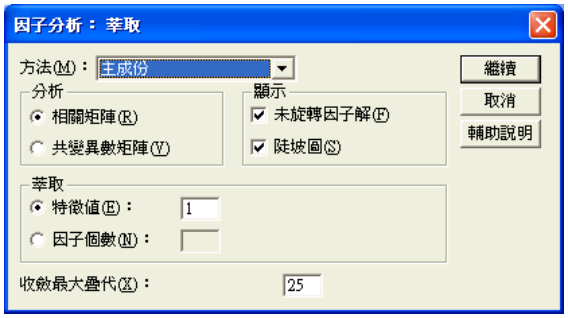
Step 2. 選擇變數



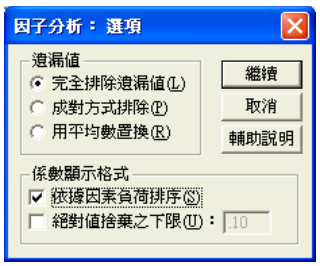
Step 3. 在『描述性統計量』方塊中，選取列出相關係數之『係數』、『行列式』、『倒數模式』（反矩陣）及『KMO與Bartlett的球形檢定』。



Step 4. 由於資料單位不同，因此爲消除單位之影響，以相關係數矩陣進行分析。『萃取』方塊中，選取輸出『陡坡圖』及特徵值大於1的主成份。



Step 5. 在『選項』方塊中，選取依據『因素負荷排序』。



由以下相關係數矩陣可發現，變數間具高度相關，因此若直接進行迴歸分析，則容易發生共線性。經Bartlett's test of sphericity之顯著值顯示P-value $= 0.000 < 0.05$，故棄卻虛無假設，即相關係數矩陣不爲一單位矩陣，故可考慮進行主成份分析。

相關矩陣

		排氣量	馬力	重量	加速時間	氣缸數
相關	排氣量	1.000	0.898	0.934	−0.548	0.951
	馬力	0.898	1.000	0.863	−0.701	0.842
	重量	0.934	0.863	1.000	−0.425	0.897
	加速時間	−0.548	−0.701	−0.425	1.000	−0.511
	氣缸數	0.951	0.842	0.897	−0.511	1.000

[a]行列式= 0.0008412

相關矩陣轉換（相關係數矩陣之反矩陣）

	排氣量	馬力	重量	加速時間	氣缸數
排氣量	20.019	−3.933	−6.219	0.526	−9.874
馬力	−3.933	9.273	−4.014	3.247	1.189
重量	−6.219	−4.014	10.319	−2.457	−1.218
加速時間	0.526	3.247	−2.457	2.698	0.347
氣缸數	−9.874	1.189	−1.218	0.347	10.658

KMO與Bartlett檢定

Kaiser-Meyer-Olkin 取樣適切性量數		0.795
Bartlett 球形檢定	近似卡方分配	2743.779
	自由度	10
	顯著性	0.000

經計算可求得相關係數矩陣之特徵值與特徵向量如下。在『解說總變異量表』中可發現第一個主成份解釋總變異的比例高達81.63%；若取兩個主成份則可解釋總變異的比例高達95.38%。讀者也可發現，此處特徵值與SPSS迴歸分析中貢獻性診斷所得之特徵值一致。

成份矩陣

	成份				
	1	2	3	4	5
排氣量	0.973	0.154	−0.062	0.012	−0.158
馬力	0.957	−0.107	0.221	0.153	0.029
氣缸數	0.945	0.189	−0.243	0.060	0.089
重量	0.931	0.288	0.127	−0.177	0.043
加速時間	−0.676	0.731	0.058	0.074	−0.003

解說總變異量

	平方和負荷量萃取		
成份	總和	變異數的%	累積%
1	4.082	81.630	81.630
2	0.687	13.750	95.380
3	0.131	2.628	98.008
4	0.064	1.278	99.286
5	0.036	0.714	100.000

由上述『解說總變異量表』可知，相關係數矩陣之特徵值及特徵向量可依計算公式（7-7）如下

特徵值

λ_1	λ_2	λ_3	λ_4	λ_5
4.082	0.687	0.131	0.064	0.036

特徵向量

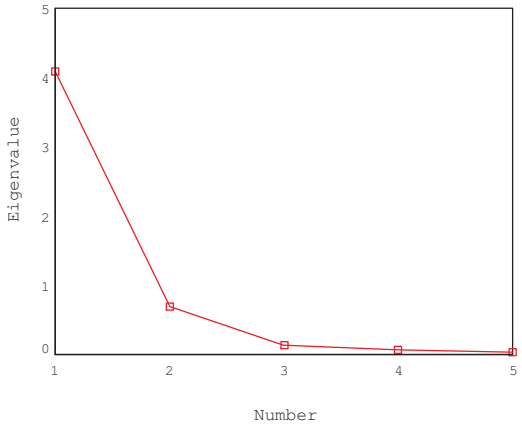
	$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$	$\mathbf{a}_4$	$\mathbf{a}_5$
排氣量	0.482	0.185	−0.172	0.049	−0.838
馬力	0.474	−0.129	0.609	0.604	0.154
氣缸數	0.461	0.347	0.352	−0.701	0.229
重量	−0.335	0.881	0.161	0.293	−0.013
加速時間	0.468	0.228	−0.671	0.238	0.471

由上述『解說總變異量表』及陡坡圖可看出，第三個以後的主成份對提高總變異的解釋能力不大，因此此例可取一至二個主成份。其中第一個主成份對每個變數的共同性如下。以排氣量爲例，選取一個主成份可解釋排氣量變異的94.7%；若取兩個主成份，怎可解釋排氣量變異的97%。很顯然地：若選取五個主成份，則原有變數資訊並不會有遺失，即對排氣量的解釋能力會有100%。但是，此種作法就喪失主成份分析縮減資料構面的精神。

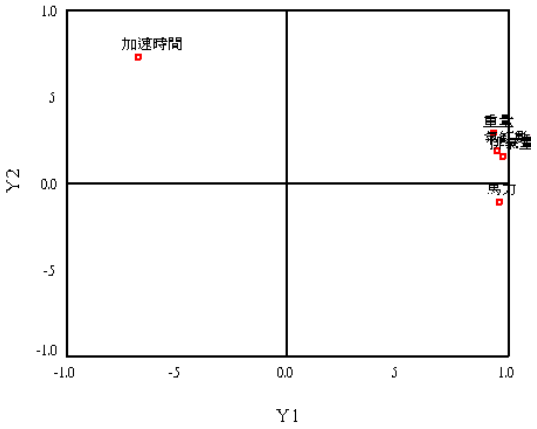
成份矩陣與變數解釋比例

	成份		共同性	
	1	2	h_1^2	h_2^2
排氣量	0.973	0.154	0.947	0.970
馬力	0.957	−0.107	0.916	0.927
氣缸數	0.945	0.189	0.893	0.929
重量	0.931	0.288	0.867	0.950
加速時間	−0.676	0.731	0.457	0.991

我們進行主成份分析之目的乃希望藉由線性轉換取得互相獨立的新變數以作爲後續研究之用。由於SPSS中所提供之主成份分析僅是因素分析中因素萃取方法之一，因此，無法由因素分析中之因素分數直接



(a) 陡坡圖



(b) 主成份負荷圖

取得主成份分數。在此例中，我們以相關係數矩陣進行分析，若想取得轉換後的主成份分數，必須先取得原始變數之標準化數值（關於變數之標準化數值，可參考前一章節說明），讀者可以下列語法檔進行運算。

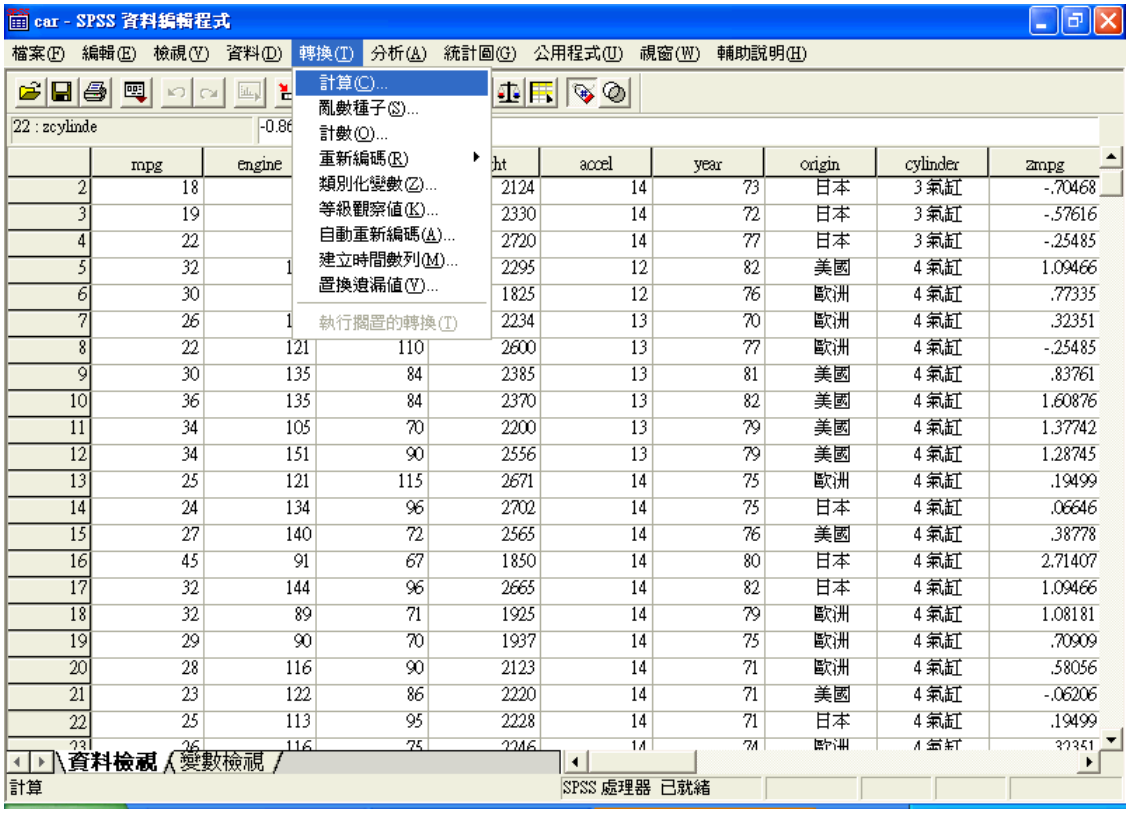
```
MATRIX.
GET X/VARIABLES=Z排氣量 Z馬力 Z重量 Z加速時間 Z氣缸數/MISSING=OMIT.
COMPUTE NR =NROW(X).
COMPUTE NC =NCOL(X).
COMPUTE XX1=SSCP(X).
COMPUTE XX2=CSUM(X).
COMPUTE XX3=XX2/NR.
 COMPUTE SIGMA=(XX1-T(XX3)*XX3*NR)/(NR-1).
COMPUTE SDIAG=MAKE(NC,NC,0).
LOOP J=1 TO NC.
COMPUTE SDIAG(J,J)=1/SQRT(SIGMA(J,J)).
END LOOP.
COMPUTE CORR=SDIAG*SIGMA*SDIAG.
PRINT SIGMA.
PRINT SDIAG.
PRINT CORR.
CALL EIGEN(SIGMA,V,LAMBDA).
PRINT LAMBDA.
PRINT V.
COMPUTE VV=T(V)*V.
PRINT VV.
COMPUTE PSCORE=X*V.
SAVE PSCORE /OUTFILE=*.
END MATRIX.
```

此外，我們也可以工具列中內建之計算功能將各別主成份分別計算其主成分分數。以第一個主成份分數爲例

$$Y_1 = 0.482 \times Z_{\text{排氣量}} + 0.474 \times Z_{\text{馬力}} + 0.461 \times Z_{\text{重量}} - 0.335 \times Z_{\text{加速時間}} + 0.468 \times Z_{\text{氣缸數}}$$

SPSS計算主成份分數方式如下:

Step 1. SPSS功能列⇒轉換⇒計算



Step 2. 將各變數與$\mathbf{a}_1$之元素輸入文字方塊。



由於各主成份間爲相互獨立，所以相關係數矩陣爲一單位矩陣，此爲必然結果。

成份分數共變數矩陣

成份	1	2	3	4	5
1	1.00	0.00	0.00	0.00	0.00
2	0.00	1.00	0.00	0.00	0.00
3	0.00	0.00	1.00	0.00	0.00
4	0.00	0.00	0.00	1.00	0.00
5	0.00	0.00	0.00	0.00	1.00

最後，當僅有樣本變異數矩陣或相關係數矩陣時，若要以SPSS進行主成份分析，則必須以以下列語法檔進行。以本章爲例語法檔如下：

```
MATRIX DATA VARIABLES = 排氣量 馬力 重量 加速時間 氣缸數
/FILE = INLINE
/FORMAT = FREE LOWER DIAG
/N = n
/CONTENTS = CORR.
BEGIN DATA.
1.000
0.898 1.000
0.934 0.863 1.000
0.548 0.701 0.425 1.000
0.951 0.842 0.897 0.511 1.000
END DATA.
FACTOR MATRIX = IN (COR = *)
/CRITERIA = FACTOR(2)
/EXTRACTION = PC
/FORMAT = SORT
/PLOT = EIGEN ROTATION(1 2).
```

第 8 章

因素分析

我們在上一章介紹到「主成份分析」，它的目的是希望用較少的變數去解釋原始資料的大部分變異。在這一章裡，我們將要介紹的因素分析法，同樣的也是希望能夠尋求一組觀測資料之精簡描述。前述主成分分析的目的再建立一組可觀測變數之線性組合使其能解釋這些變數之總變異之大部分。而因素分析則是希望在一群具有相關性且難以解釋的資料中，找出幾個概念上有意義的，並且彼此之間近於獨立的可以影響原始資料的共同因素，使盡可能再製原共變異數矩陣。

主成份分析所著重的在於如何「轉換」原始變項使之成爲一些綜合性的新指標，即將原資料做一個正交轉換，而其關鍵在「變異數」問題。與主成份分析不同的是，它是假設所看到的資料和某些隱藏的、不可觀測的「因素」之間有一線性迴歸的形式（因此，容許「誤差項」）。而因素分析的目標就是要找出這些不可見的因素及它和各觀測資料項的關係。

8.1 正交因子模式

設$\mathbf{X}$爲$p \times 1$可觀測隨機向量，其均值爲$\boldsymbol{\mu}$，共變異數矩陣爲$\boldsymbol{\Sigma}$, 則$\mathbf{X}$之m因子模式爲：

$$\begin{aligned} X_1 - \mu_1 &= \ell_{11}F_1 + \ell_{12}F_2 + \cdots + \ell_{1m}F_m + \varepsilon_1 \\ X_2 - \mu_2 &= \ell_{21}F_1 + \ell_{22}F_2 + \cdots + \ell_{2m}F_m + \varepsilon_2 \\ \vdots & \qquad\qquad \vdots \\ X_p - \mu_p &= \ell_{p1}F_1 + \ell_{p2}F_2 + \cdots + \ell_{pm}F_m + \varepsilon_p \end{aligned}$$

若以矩陣形式表示, 則爲

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{LF} + \boldsymbol{\varepsilon} \tag{8-1}$$

其中$\mathbf{L}(p \times m)$爲因素負荷矩陣 (matrix of factor loadings)，元素ℓ_{ij}是第i個變數在第j個因子F_j之因素負荷量 (factor loading)，爲每一因子之權數, 其平方表示X_i之變異數由F_j解釋的部分。$\mathbf{L}$又稱爲因子組型矩

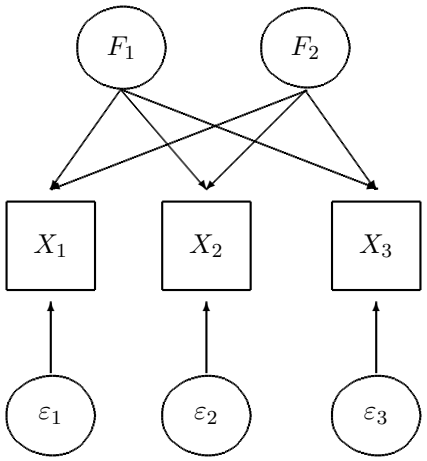


圖 8.1: 因素分析示意圖

陣 (factor pattern matrix)。$\mathbf{F}(m \times 1)$爲不可觀測隨機向量，其元素$\mathbf{F_1}, \cdots, \mathbf{F_m}$稱爲共同因子 (common factors)，而$\boldsymbol{\varepsilon}(p \times 1)$爲隨機誤差向量，有時稱之爲特定 (獨特) 因子 (specific or unique factors)，其第i個元素ε_i僅屬於X_i。對於此模式我們假設

$$\begin{aligned} &\mathrm{E}(\mathbf{F}) = \mathbf{0}, \\ &\mathrm{Cov}(\mathbf{F}) = \mathrm{E}[(\mathbf{F} - \mathrm{E}(\mathbf{F}))(\mathbf{F} - \mathrm{E}(\mathbf{F}))^\top] = \mathrm{E}(\mathbf{F}\mathbf{F}^\top) = \mathbf{I}, \\ &\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}, \\ &\mathrm{Cov}(\boldsymbol{\varepsilon}) = E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top] = \boldsymbol{\Psi} = \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix} \end{aligned} \tag{8-2}$$

以及$\mathbf{F}$與$\boldsymbol{\varepsilon}$是獨立的，即

$$\mathrm{Cov}(\boldsymbol{\varepsilon}, \mathbf{F}) = \mathrm{E}(\boldsymbol{\varepsilon}\mathbf{F}^\top) = \mathbf{0} \tag{8-3}$$

因此，所有因子彼此間是無關的，而且共同因子都標準化，其均值都爲 0，變異數都爲 1。 在這些假定下 (8-1) 式即爲正交因子模式 (orthogonal factor model)。 (8-2) 式中之ψ_i爲X_i之變異數$V(X_i) = \sigma_{ii}$中

屬於特定因子的部分稱爲特定變異數 (specific variance) 或獨特性 (uniquenes)。在正交因子模式下，$\mathbf{X}$之共變異數矩陣$\mathbf{\Sigma}$可寫成

$$\begin{aligned}\mathbf{\Sigma} &= \mathrm{Cov}(\mathbf{X}) \\ &= \mathrm{E}(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^{\top} \\ &= \mathrm{E}(\mathbf{LF}+\boldsymbol{\varepsilon})(\mathbf{LF}+\boldsymbol{\varepsilon})^{\top} \\ &= \mathbf{L}\mathbf{L}^{\top}+\mathbf{\Psi} \end{aligned} \tag{8-4}$$

故可知

$$Var(X_i) = \ell_{i1}^2 + \cdots + \ell_{im}^2 + \psi_i = \sum_{j=1}^{m} \ell_{ij}^2 + \psi_i = h_i^2 + \psi_i$$

另外，$\mathbf{X}$與$\mathbf{F}$之共變異數爲

$$\begin{aligned}\mathrm{Cov}(\mathbf{X},\mathbf{F}) &= \mathrm{E}(\mathbf{X}-\boldsymbol{\mu})\mathbf{F}^{\top} \\ &= \mathrm{E}\left[(\mathbf{LF}+\boldsymbol{\varepsilon})\mathbf{F}^{\top}\right] \\ &= \mathbf{L}\mathrm{E}\left(\mathbf{F}\mathbf{F}^{\top}+\boldsymbol{\varepsilon}\mathbf{F}^{\top}\right) \\ &= \mathbf{L} \end{aligned} \tag{8-5}$$

由（8-5）可知，第i個變數X_i與第j個因素之共變異數爲$\mathrm{Cov}(X_i, F_j) = \ell_{ij}$。且由 (8-4) 知任一$X_i$之變異數$\sigma_{ii}$等於$\sum_{j=1}^{m} \ell_{ij}^2 + \psi_i$，其中$\sum_{j=1}^{m} \ell_{ij}^2$可視爲$X_i$的變異數中，由$m$個共同因子所能解釋的部分，稱之爲$X_i$的共同性（communality），以$h_i^2$表示之。而無法由$m$個因子所解釋的部份，則稱之爲獨特性（uniquenes），以ψ_i表示之。其次, 由 (8-5)式可知, 在正交因子模式中，因子負荷ℓ_{ij}爲X_i與F_j之共變異數, 當$\mathbf{X}$之變異數爲1時, 則ℓ_{ij}爲X_i與 F_j之相關係數。因此，$\mathbf{L}$爲$\mathbf{X}$與$\mathbf{F}$之相關矩陣，又稱因子結構矩陣（factor structure matrix）。

8.2 因素負荷估計方法

一般常用於m因子模式估計參數的方法有主成分分析法 (principal components analysis，PCA)，主因子分析法 (principal factor analysis，PFA)，最大概似法 (maximum likelihood method)，茲介紹如下：

1 主成份分析法：目的在找出原變項的線性組合，使得這些主成份的變異量爲最大。

2 主因子分析法：將估計的共同性（起始值通常爲SMC，squared multiple correlations）取代$\mathbf{R}$之對角線元素開始，從萃取的因子負荷估計共同性，再自此新共同性取代$\mathbf{R}$之對角線元素，重估共通性，逐次疊代修正至共通性收斂至一固定值爲止。

3 最大概似法：假定樣本來自多變量常態母體，此法即尋找最可能產生觀測的相關係數矩陣的參數估計值，$\mathbf{L}$和$\mathbf{\Psi}$使得對數概似函數

$$-\frac{1}{2}(n-1)\log(|\mathbf{\Sigma}|)-\frac{1}{2}(n-1)\text{trace}\left(\mathbf{S}\mathbf{\Sigma}^{-1}\right)+\text{constant}$$

有極大值，其中n爲樣本數及$\mathbf{S}$爲樣本相關係數矩陣。

8.3 共同因素個數的選取

因素分析的主要目的在將繁多的變項縮減爲少數的因素，找出變項背後的結構。然而，到底需要幾個因素才足夠，這沒有一定的標準答案。因素的個數選可參考前一章主成份分析所提供之方式。若資料爲抽自多變量母體時，也可利用最大相似比檢定（likelihood ratio test）來檢定適當的因素個數。假設當k個因素被選取時，檢定其餘的$p-k$個特徵值是否相等，檢定統計量爲

$$\chi^2=\left(n-\frac{2p+5}{6}\right)\left[-\sum_{i=k+1}^{p}\log\lambda_i+(p-k)\log\left(\sum_{i=k+1}^{p}\frac{\lambda_i}{p-k}\right)\right] \tag{8-6}$$

其中，$\chi^2\sim\chi^2\left(\frac{1}{2}(p-k-1)(p-k+2)\right)$。所以，當$k=0$時，即爲檢定所有特徵值是否均相等。若不棄卻虛無假設，則表示所有因素之貢獻均爲相等，因此不需作因素分析。此一檢定爲連續檢定$k=0,1,2,\ldots$直到棄卻虛無假設爲止。接著可以模式適合度檢定（test the goodness of fit）檢定因素個數爲k時，模式是否適當。其檢定統計量爲

$$\chi^2=\left[n-1-\frac{2p+5}{6}-\frac{2k}{3}\right]\left[\sum_{j=k+1}^{p}(\lambda_j-\log\lambda_j)-(p-k)\right] \tag{8-7}$$

其中，$\chi^2\sim\chi^2\left(\frac{1}{2}[(p-k)^2-(p+k)]\right)$。必須注意的是，該檢定對樣本數高度敏感，因此常會導致棄卻虛無假設$H_0:\mathbf{\Sigma}=\mathbf{L}\mathbf{L}^\top+\mathbf{\Psi}$。

此外，KMO量測值（KMO，Kaiser-Meyer-Olkin measure of sampling adequacy）及前一章所提之Bartlett球形檢定（7.2）可以用來幫助判斷資料是否適合進行因素分析。KMO量測值定義如下：

$$\text{KMO} = \frac{\sum_i^p \sum_{j \neq i} r_{ij}^2}{\sum_i^p \sum_{j \neq i} r_{ij}^2 + \sum_i^p \sum_{j \neq i} s_{ij}^2} \tag{8-8}$$

KMO量測值數值介於0與1之間，當偏相關係數越小（即表示該兩變數無法被變數所取代），則數值愈靠近1，表示變項的相關愈高，愈適合進行因素分析；數值愈靠近0，表示變項的相關愈低，愈不適合進行因素分析。一般KMO量測值的評判標準如下表：

KMO	適當性
> 0.90	極佳
$0.80 \sim .89$	良好
$0.70 \sim .79$	中度
$0.60 \sim .69$	平庸
$0.50 \sim .59$	粗劣
< 0.50	無法接受

8.4 因素轉軸

由於因素解非唯一，故當原始的因素負荷矩陣不易解釋時，可將因素作旋轉。進行轉軸之後，會導致變項在每個因素的負荷量變大或變小，目的在將因素負荷量調整成接近1或0的數值，以更突顯變項與因素之關係。因素轉軸的方法分爲「直交轉軸」（orthogonal rotations）和「斜交轉軸」（oblique rotations）兩種。在直交轉軸法中，因素與因素之間沒有相關，因素軸之間的夾角等於90度，即表示轉軸後因素與因素之間仍是相互獨立。斜交轉軸後，因素軸之間的夾角並非90度，表示因素與因素之間彼此有某種程度的相關。一般使用以直交轉軸法爲主，其轉換方式有

1 最大變異法（Varimax）：尋找一因素結構使得因素負荷量的變異數在因素內最大。

2 四次方最大值法（Quartimax）：尋找一因素結構使得因素負荷量的變異數在變項內最大。

3 相等最大值法（Equamax）：綜合前兩者，尋找一因素結構使得因素負荷量的變異數在因素內與變項內同時最大。

8.5 因素分數

我們進行因素分析之目的乃希望藉由共同因素的取得以作爲後續研究之用。因素分數的取得與主成份分數

不同，主要以因素負荷量爲基礎，透過迴歸原理來獲得一組因素分數係數。因素分數計算的方法一般常用的有：

1 普通最小平方法：求解最適$\mathbf{F}$使得誤差平方和極小化爲目標。由（8-1）可知， $\boldsymbol{\varepsilon}^{\top}\boldsymbol{\varepsilon} = (\mathbf{X}-\boldsymbol{\mu}-\mathbf{LF})^{\top}(\mathbf{X}-\boldsymbol{\mu}-\mathbf{LF})$，因此可求得極小化必要條件爲

$$\mathbf{L}^{\top}(\mathbf{X}-\boldsymbol{\mu}) = \mathbf{L}^{\top}\mathbf{LF}$$

移項後可推得

$$\mathbf{F} = (\mathbf{L}^{\top}\mathbf{L})^{-1}\mathbf{L}^{\top}(\mathbf{X}-\boldsymbol{\mu}) \tag{8-9}$$

由於母體平均數未知，因此以樣本平均數矩陣 $\overline{\mathbf{X}}$作爲估計值，所以

$$\widehat{\mathbf{F}} = \left(\widehat{\mathbf{L}}^{\top}\widehat{\mathbf{L}}\right)^{-1}\widehat{\mathbf{L}}^{\top}\left(\mathbf{X}-\overline{\mathbf{X}}\right) \tag{8-10}$$

2 加權最小平方法：由Bartlett於1937年所提出，以$\boldsymbol{\Psi}$爲權數，求解最適$\mathbf{F}$使得加權誤差平方和極小化爲目標。由（8-1）可知， $\boldsymbol{\varepsilon}^{\top}\boldsymbol{\Psi}^{-1}\boldsymbol{\varepsilon} = (\mathbf{X}-\boldsymbol{\mu}-\mathbf{LF})\boldsymbol{\Psi}^{-1}(\mathbf{X}-\boldsymbol{\mu}-\mathbf{LF})^{\top}$，因此可求得極小化必要條件爲 $\mathbf{L}^{\top}\boldsymbol{\Psi}^{-1}(\mathbf{X}-\boldsymbol{\mu}) = \mathbf{L}^{\top}\boldsymbol{\Psi}^{-1}\mathbf{LF}$由於母體平均數，因此分別以樣本平均數矩陣及殘差矩陣$\overline{\mathbf{X}}$及$\widehat{\boldsymbol{\Psi}}$作爲估計值。所以

$$\widehat{\mathbf{F}} = \left(\widehat{\mathbf{L}}^{\top}\widehat{\boldsymbol{\Psi}}^{-1}\widehat{\mathbf{L}}\right)^{-1}\widehat{\mathbf{L}}^{\top}\widehat{\boldsymbol{\Psi}}^{-1}\left(\mathbf{X}-\overline{\mathbf{X}}\right) \tag{8-11}$$

3 迴歸法：由Thomson於1958年所提出，若以最大概似估計法進行因素萃取時，以未知的F對X作迴歸估計。由於Cove$(\mathbf{X},\mathbf{F}) = \mathbf{L}$，故可知

$$\mathrm{E}(\mathbf{F}|\mathbf{X}=\mathbf{x}) = \mathrm{E}(\mathbf{F}) + \mathbf{L}^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu}) \tag{8-12}$$

及

$$\mathrm{Cov}(\mathbf{F}|\mathbf{X}=\mathbf{x}) = \mathbf{I} - \mathbf{L}^{\top}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{L} \tag{8-13}$$

又母體均數及變異數未知及$\mathrm{E}(\mathbf{F}) = \mathbf{0}$，由（8-5），故可推得

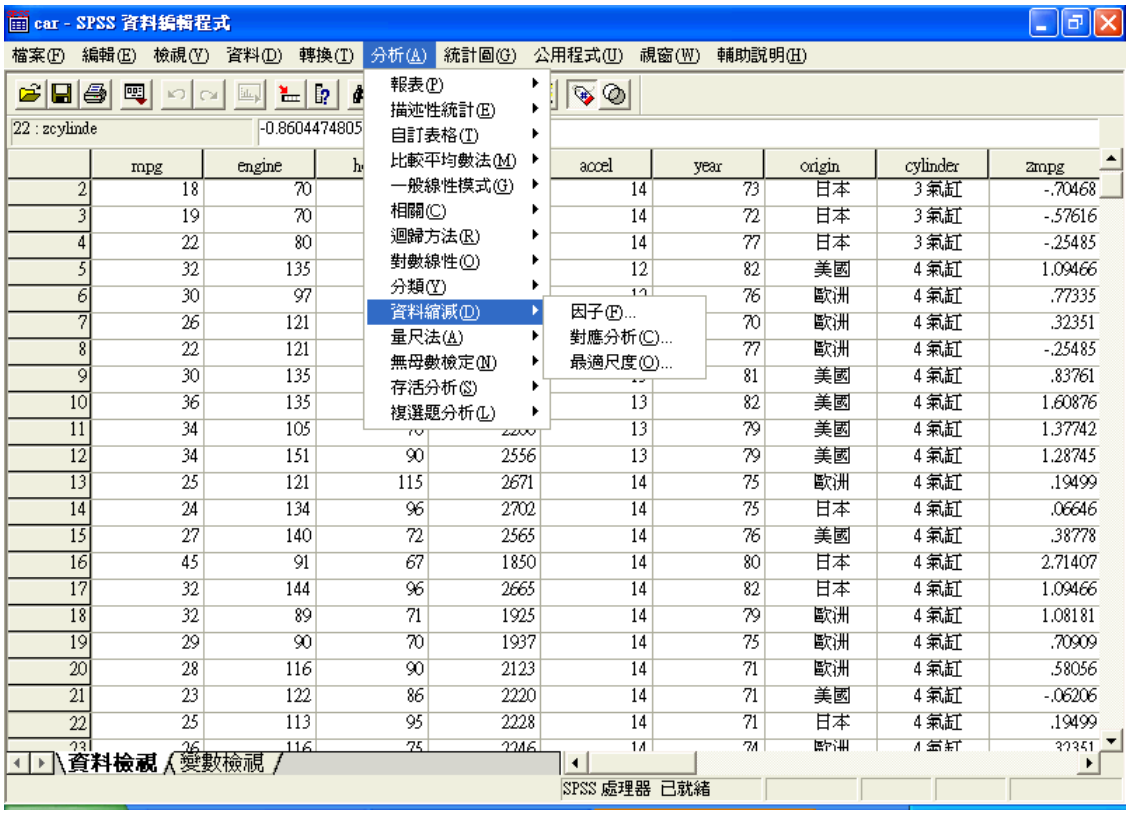
$$\widehat{\mathbf{F}} = \widehat{\mathbf{L}}^{\top}\mathbf{S}^{-1}(\mathbf{X}-\overline{\mathbf{X}}) \tag{8-14}$$

4 Anderson-Rubin法：由Anderson和Rubin於1956年所提出，修正加權最小平方法以確保因素間是正交性。因此，產生的因素分數平均數爲0， 變異數爲1且彼此不相關。

8.6 範例

以SPSS內部檔案cars.sav爲例，試以排氣量、馬力、重量、加速時間及氣缸數進行因素分析。

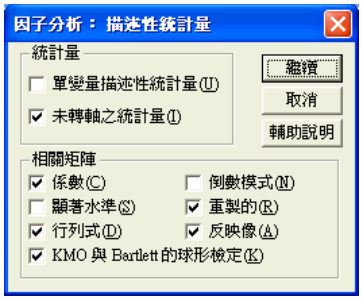
Step 1. SPSS功能列⇒分析⇒資料縮減⇒因子



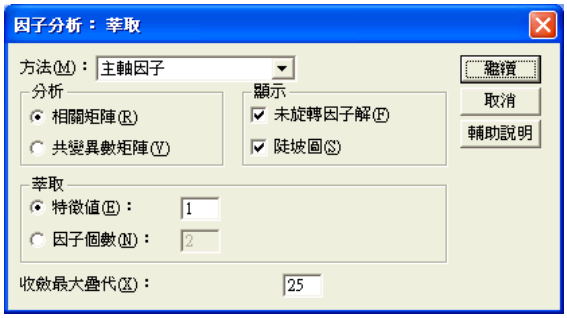
Step 2. 選擇變數



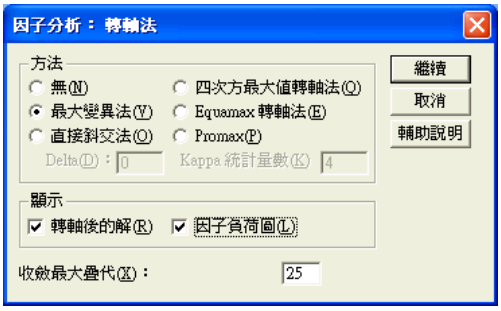
Step 3. 在『描述性統計量』方塊中，選取列出相關係數之『係數』、『行列式』、『重製的』、『反映像』（偏相關係數矩陣）及『KMO與Bartlett的球形檢定』。



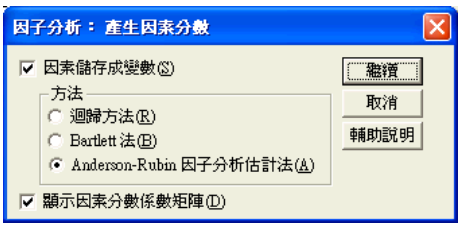
Step 4. 由於資料單位不同，因此爲消除單位之影響，以相關係數矩陣進行分析。『萃取』方法選擇主軸因子（主因子分析法），選取輸出『陡坡圖』及特徵值大於1的因素。



Step 5. 在『轉軸』方法中，首先選擇一般常用的『最大變異法』及輸出『因子負荷圖』。



Step 6. 在『分數』中，以Anderson-Rubin法取得因素分數並選擇『因素儲存成變數』。



SPSS報表輸出如下：

由於反映像矩陣內偏相關值不高，表示資料適合進行因素分析。 且由『KMO 與 Bartlett 檢定』表可看出，KMO值爲0.795，Bartlett's球形考驗近似卡方值之P-value=0，顯示有隱藏因素存在。

反映像矩陣

		排氣量	馬力	重量	加速時間	氣缸數
反映像相關	排氣量	0.797	−0.289	−0.433	0.072	−0.676
	馬力	−0.289	0.800	−0.410	0.649	0.120
	重量	−0.433	−0.410	0.816	−0.466	−0.116
	加速時間	0.072	0.649	−0.466	0.656	0.065
	氣缸數	−0.676	0.120	−0.116	0.065	0.846

KMO與Bartlett檢定

Kaiser-Meyer-Olkin 取樣適切性量數。		0.795
Bartlett 球形檢定	近似卡方分配	2743.779
	自由度	10
	顯著性	0.000

由於僅有一特徵值大於1，因此在本例僅取一因素。藉由主軸因子法，經由6次迭代運算可求得因素負荷$\hat{\mathbf{L}}$及共同性如下：

主軸因子萃取法迭代收斂表

		第一次	第二次	第三次	第四次	第五次	第六次
	特徵值	3.9654	3.9333	3.9302	3.9307	3.9312	3.93144
排氣量	因素負荷	0.9764	0.9816	0.9845	0.9859	0.9865	0.9867
馬力		0.9440	0.9462	0.9472	0.9477	0.9478	0.9479
重量		0.9249	0.9187	0.9164	0.9155	0.9151	0.9149
加速時間		−0.6206	−.5848	−.5785	−.5774	−.5772	−.5772
氣缸數		0.9383	0.9426	0.9430	0.9430	0.9429	0.9429
排氣量	共同性	0.9534	0.9635	0.9692	0.9719	0.9731	0.9736
馬力		0.8911	0.8953	0.8973	0.8981	0.8984	0.8985
重量		0.8554	0.8439	0.8398	0.8381	0.8374	0.8371
加速時間		0.3851	0.3420	0.3347	0.3334	0.3332	0.3331
氣缸數		0.8804	0.8885	0.8893	0.8892	0.8891	0.8890

因素個素：1

因子矩陣

	因子 1
排氣量	0.988
馬力	0.949
重量	0.916
加速時間	−0.578
氣缸數	0.937

萃取方法：主軸因子。

由因子矩陣表可推得共同性，及獨特性，其中初始值即爲SMC。由下表可發現以主軸法取一因素可解釋97.6%之排氣量、90.1%之馬力、89.3%重量、33.4%加速時間及87.7%氣缸數。

因素負荷、共同性及獨特性表

	初始	因素負荷	萃取（共同性）	獨特性
排氣量	0.950	0.988	0.976	0.024
馬力	0.892	0.949	0.901	0.099
重量	0.903	0.916	0.839	0.169
加速時間	0.629	−0.578	0.334	0.666
氣缸數	0.906	0.937	0.877	0.123

萃取法：主軸因子萃取法。

此外，經主軸因子萃取法所萃取之因素能解釋總體變異之78.529%。最後，藉由$\widehat{\mathbf{L}}\widehat{\mathbf{L}}^{\top}$求得重製相關矩陣，其中殘差矩陣爲$\mathbf{R}-\widehat{\mathbf{L}}\widehat{\mathbf{L}}^{\top}-\widehat{\mathbf{\Psi}}$。

解說總變異量

初始特徵值		平方和負荷量萃取				
因子	總和	變異數的%	累積%	總和	變異數的%	累積%
1	4.082	81.630	81.630	3.926	78.529	78.529
2	0.687	13.750	95.380			
3	0.131	2.628	98.008			
4	0.064	1.278	99.286			
5	0.036	0.714	100.000			

萃取法：主軸因子萃取法。

重製矩陣

		排氣量	馬力	重量	加速時間	氣缸數
重製相關	排氣量	0.976	0.938	0.905	−0.571	0.925
	馬力	0.938	0.901	0.869	−0.548	0.889
	重量	0.905	0.869	0.839	−0.529	0.858
	加速時間	−0.571	−0.548	−0.529	0.334	−0.541
	氣缸數	0.925	0.889	0.858	−0.541	0.877
殘差	排氣量		−.0392	0.0291	0.0224	0.0257
	馬力	−0.0392		−0.0064	−0.1530	−0.0466
	重量	0.0291	−.0064		0.1036	0.0394
	加速時間	0.0224	−0.1530	0.1036		0.0301
	氣缸數	0.0257	−0.0466	0.0394	0.0301	

若以最大概似法萃取2因素，則經由最大變異法轉軸後所估計之$\widehat{\mathbf{L}}$如下表，可發現兩因素可解釋98.4%之排氣量、88.2%之馬力、90.4%重量、62.9%加速時間及90.6%氣缸數。

因子矩陣

	因子	
	1	2
排氣量	0.941	0.314
馬力	0.785	0.515
重量	0.932	0.190
加速時間	−0.260	−0.965
氣缸數	0.913	0.283

共同性及獨特性

	初始	萃取（共同性）	獨特性
排氣量	0.950	0.984	0.016
馬力	0.892	0.882	0.118
重量	0.903	0.904	0.096
加速時間	0.629	0.999	0.001
氣缸數	0.906	0.913	0.087

萃取法：最大概似。

萃取兩個因素之模式，經由卡方適合度檢定顯著性 ，因此棄卻虛無假設，即模式不合適。如同之前介紹，該檢定對樣本數高度敏感，因此常會導致棄卻虛無假設。此外，藉由$\widehat{\mathbf{L}}\widehat{\mathbf{L}}^\top$求得重製相關矩陣，其中殘差矩陣爲$\mathbf{R}-\widehat{\mathbf{L}}\widehat{\mathbf{L}}^\top-\widehat{\mathbf{\Psi}}$。由殘差矩陣可發現樣本相關係數矩陣與重製相關矩陣之間的殘差並無絕對值>0.05的非重複性殘差，所以可知模式適合度很高。

適合度檢定

卡方	自由度	顯著性
60.346	1	0.000

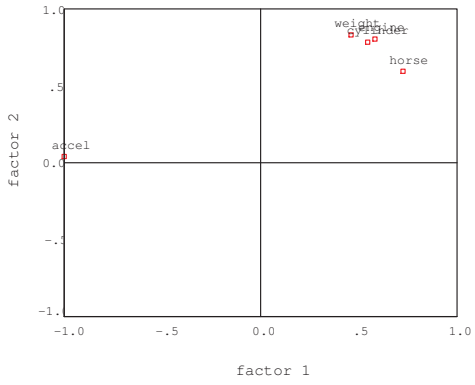
重製矩陣

		排氣量	馬力	重量	加速時間	氣缸數
重製相關	排氣量	0.984	0.901	0.936	−0.548	0.948
	馬力	0.901	0.882	0.829	−0.701	0.863
	重量	0.936	0.829	0.904	−0.426	0.904
	加速時間	−0.548	−0.701	−0.426	0.999	−0.511
	氣缸數	0.948	0.863	0.904	−0.511	0.913
殘差	排氣量		−0.002	−0.002	0.000	0.003
	馬力	−0.002		0.034	0.000	−0.020
	重量	−0.002	0.034		0.000	−0.007
	加速時間	0.000	0.000	0.000		0.000
	氣缸數	0.003	−0.020	−0.007	0.000	

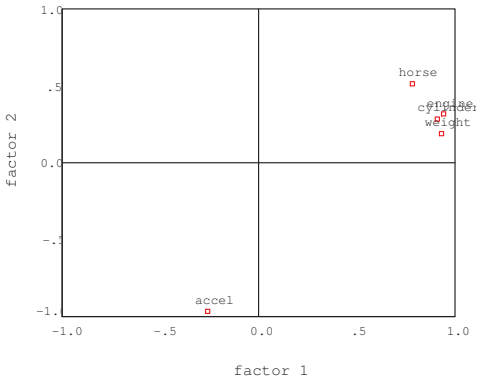
最後，由下表及轉軸前後之因素負荷圖可發現，因素轉軸前後各因素的解釋能力雖然會改變，但共同性及整體模式的解釋能力並不會隨轉軸而改變。換句話說，因素的轉軸僅是爲了方便因素的解釋或命名，並使各個因素的意義變得比較清晰明顯，並不會改變模式的解釋能力。

不同轉軸方法因素負荷量

	未轉軸			最大變異法			四次方最大值法		
變項	F_1	F_2	h^2	F_1	F_2	h^2	F_1	F_2	h^2
排氣量	0.973	0.154	0.971	0.930	0.325	0.971	0.984	−0.05	0.971
馬力	0.931	0.288	0.951	0.957	0.187	0.951	0.971	0.09	0.951
重量	0.957	−0.107	0.927	0.793	0.547	0.927	0.914	−0.301	0.927
加速時間	−0.676	0.731	0.991	−0.251	−0.963	0.991	−0.512	0.854	0.991
氣缸數	0.945	0.189	0.929	0.922	0.280	0.929	0.964	−0.009	0.929
變異數%	81.7%	13.7%	95.4%	66.5%	28.9%	95.4%	78.8%	16.6%	95.4%



(a) 轉軸前



(b) 轉軸前

第 9 章

典型相關分析

當研究的變項只有X、Y兩個變項時，這兩個變項的線性相關爲簡單相關；當研究變項只有一個Y變項、但卻有p個X變項時，這p個X變項與一個Y變項之間的線性相關爲多元相關；當研究變項同時有p個X變項、q個Y變項時，這p個X變項與q個Y變項之間的線性相關爲典型相關。

典型相關分析（Canonical Correlation Analysis）爲Hotelling於1935年所提出，用以分析兩組變數間關係的一種分析方法。它所描述的是兩組變數組間的整體相關形式，而不是關於兩組變數中個別變數的相關。其原理爲藉由將兩組變數進行直線組合，使之成爲兩個新變數，此兩個新的變數稱爲典型變量（canonical variables）。根據這兩個新變數計算得到的相關係數稱爲典型相關係數（canonical correlations）。

9.1 典型相關模式

在典型相關分析中，有以下三個主要問題：

1 如何決定各組組內的變數在線性組合的權重值？

2 典型變數間的相互關係？

3 以及典型變數與原來變數間的關係如何？

設$\mathbf{X}$、$\mathbf{Y}$分別爲$p \times 1$及$q \times 1$的可觀測隨機向量，其中

$$\begin{aligned} \mathrm{E}\mathbf{X} &= \boldsymbol{\mu}_1, \quad \mathrm{Cov}(\mathbf{X}) = \boldsymbol{\Sigma}_{11} \\ \mathrm{E}\mathbf{Y} &= \boldsymbol{\mu}_2, \quad \mathrm{Cov}(\mathbf{Y}) = \boldsymbol{\Sigma}_{22} \\ \mathrm{Cov}(\mathbf{X}, \mathbf{Y}) &= \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^{\top} \end{aligned} \tag{9-1}$$

爲便於說明起見，令$p \le q$。兩組變數$\mathbf{X}$與$\mathbf{Y}$，各自經過線性組合後成爲典型變數：

$$U_1 = \mathbf{a}_1^\top \mathbf{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \tag{9-2}$$

及

$$V_1 = \mathbf{b}_1^\top \mathbf{Y} = b_{11}Y_1 + b_{12}Y_2 + \cdots + b_{1q}Y_q \tag{9-3}$$

典型相關分析的目的在於找出使p個X變項之線性組合與q個Y變項之線性組合之相關達到最大值，即

$$\max \quad \mathrm{Corr}(U_1, V_1) = \frac{\mathbf{a}_1^\top \boldsymbol{\Sigma}_{12}\mathbf{b}_1}{\sqrt{\mathbf{a}_1^\top \boldsymbol{\Sigma}_{11}\mathbf{a}_1}\sqrt{\mathbf{b}_1^\top \boldsymbol{\Sigma}_{22}\mathbf{b}_1}} \tag{9-4}$$

爲防止求解的過程中兩向量的值趨向無限大，我們有以下假設

$$\mathbf{a}_1^\top \boldsymbol{\Sigma}_{11}\mathbf{a}_1 = 1, \mathbf{b}_1^\top \boldsymbol{\Sigma}_{22}\mathbf{b}_1 = 1$$

典型相關分析建立第一個典型變量的原則是使所建立的兩個典型變量之間的相關係數最大化。 然後，繼續在兩組變數剩餘的變化中尋找第二個最大的相關係數，此爲第二對典型變量。重複此一步驟，最多可以導出$\min(p, q)$個典型相關係數。

假設$\mathbf{e}_1 = \boldsymbol{\Sigma}_{11}^{1/2}\mathbf{a}_1$及$\mathbf{f}_1 = \boldsymbol{\Sigma}_{22}^{1/2}\mathbf{b}_1$，則（9-4）可改寫爲

$$\frac{\mathbf{a}_1^\top \boldsymbol{\Sigma}_{12}\mathbf{b}_1}{\sqrt{\mathbf{a}_1^\top \boldsymbol{\Sigma}_{11}\mathbf{a}_1}\sqrt{\mathbf{b}_1^\top \boldsymbol{\Sigma}_{22}\mathbf{b}_1}} = \frac{\mathbf{e}_1^\top \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}\mathbf{f}_1}{\sqrt{\mathbf{e}_1^\top \mathbf{e}_1}\sqrt{\mathbf{f}_1^\top \mathbf{f}_1}} \tag{9-5}$$

由科西不等式可知

$$\begin{aligned}
& \left[\mathbf{e}_1^\top \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}\right]\mathbf{f}_1 \\
\le\ & \left\{\left[\mathbf{e}_1^\top \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}\right]\left[\mathbf{e}_1^\top \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}\right]^\top\right\}^{1/2}\left\{\mathbf{f}_1^\top \mathbf{f}_1\right\}^{1/2} \\
=\ & \left\{\mathbf{e}_1^\top \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}\mathbf{e}_1\right\}^{1/2}\left\{\mathbf{f}_1^\top \mathbf{f}_1\right\}^{1/2} \\
=\ & \left\{\mathbf{e}_1^\top \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}\mathbf{e}_1\right\}^{1/2}\left\{\mathbf{f}_1^\top \mathbf{f}_1\right\}^{1/2}
\end{aligned}$$

因此目標函數可改寫爲

$$\max \quad \mathbf{e}_1^\top \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}\mathbf{e}_1$$

令 $L = \mathbf{e}_1^\top \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}\mathbf{e}_1 - \lambda\left(\mathbf{e}_1^\top\mathbf{e}_1 - 1\right)$，則最佳解的必要條件爲

$$\frac{\partial L}{\partial \mathbf{e}_1} = 2\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}\mathbf{e}_1 - 2\lambda\mathbf{e}_1 = 0$$

移項整理後可得

$$\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}\mathbf{e}_1 = \lambda_1\mathbf{e}_1$$

故

$$\max \quad \frac{\mathbf{e}_1^\top \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}\mathbf{e}_1}{\mathbf{e}_1^\top\mathbf{e}_1} = \lambda_1$$

其中λ_1及$\mathbf{e}_1$分別爲$\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}$之最大特徵根及特徵向量。將此一結果代入（9-5）可得

$$\max \quad \frac{\mathbf{e}_1^\top \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}\mathbf{f}_1}{\sqrt{\mathbf{e}_1^\top\mathbf{e}_1}\sqrt{\mathbf{f}_1^\top\mathbf{f}_1}} = \lambda_1$$

故可推得第一對典型變量之相關係數爲

$$\max \quad \mathrm{Corr}(\mathbf{a}_1^\top\mathbf{X}, \mathbf{b}_1^\top\mathbf{Y}) = \sqrt{\lambda_1} \tag{9-6}$$

其中，λ_1爲$\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}$之最大特徵根；且 $\mathbf{a}_1 = \boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\mathbf{e}_1$ 及 $\mathbf{b}_1 = \boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}\mathbf{f}_1$。我們稱此線性組合的係數爲典型權重（canonical weight）；U_i與V_i之相關係數爲典型相關係數（canonical correlations）。由於$\mathbf{e}_1$爲λ_1所對應之特徵向量，由特徵向量之定義可知

$$\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\mathbf{e}_1 = \lambda_1\mathbf{e}_1 \tag{9-7}$$

上式左右兩邊同乘$\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}$，可得

$$\left(\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}\right)\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\mathbf{e}_1 = \lambda_1\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\mathbf{e}_1 \tag{9-8}$$

故可求得

$$\mathbf{f}_1 = \boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\mathbf{e}_1 \tag{9-9}$$

爲$\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}$之最大特徵值$\lambda_1$所對應之特徵向量。再者，將（9-7）左右兩邊同乘$\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}$，可推得

$$\left(\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\mathbf{e}_1 = \lambda_1\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\mathbf{e}_1 \tag{9-10}$$

因此，求解$\mathbf{a}_1$、$\mathbf{b}_1$亦可藉由求解$\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$與$\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$之特徵向量著手。但必須注意的是，以上兩矩陣皆爲非對稱矩陣。同理，$\mathbf{a}_2$、$\mathbf{b}_2$即爲第二大的特徵值λ_2所謂應的特徵向量；$\mathbf{a}_k$、$\mathbf{b}_k$，$k = 1, 2, \ldots \min(p, q)$則可類推。

由於$\mathbf{e}_i$、$\mathbf{f}_i$皆爲特徵向量，典型變量（U_i、V_i）有下列性質：

1 $\mathrm{Var}(U_i) = \mathrm{Var}(V_i) = 1$

2 $\mathrm{Cov}(U_i, U_j) = 0 \quad \text{for} \quad i \neq j$

3 $\mathrm{Cov}(V_i, V_j) = 0 \quad \text{for} \quad i \neq j$

4 $\mathrm{Cov}(U_i, V_j) = 0 \quad \text{for} \quad i \neq j$

性質4成立的原因爲$\mathbf{f}_j = \mathbf{\Sigma}_{22}^{-\frac{1}{2}}\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-\frac{1}{2}}\mathbf{e}_j$，又$\mathbf{e}_i \perp \mathbf{e}_j$，故可得$\mathrm{Cov}(U_i, V_j) = 0$。此外，典型變量（$U_i$、$V_i$）與各組變數之典型相關係數爲

1 $\mathrm{Corr}(U_i, \mathbf{X}) = \mathbf{a}_i^\top \mathbf{\Sigma}_{11}\mathbf{V}_{11}^{-\frac{1}{2}}$，$\mathbf{V}_{11}$爲對角線爲$\mathrm{Var}(X_i)$之矩陣

2 $\mathrm{Corr}(V_j, \mathbf{Y}) = \mathbf{b}_j^\top \mathbf{\Sigma}_{22}\mathbf{V}_{22}^{-\frac{1}{2}}$，$\mathbf{V}_{22}$爲對角線爲$\mathrm{Var}(Y_j)$之矩陣

我們稱此一相關係數爲典型負荷量（canonical loadings）。若$\mathbf{X}$與$\mathbf{Y}$皆爲標準化資料，則上述性質可改寫爲

1 $\mathrm{Corr}(U_i, \mathbf{X}) = \mathbf{a}_i^\top \mathbf{R}_{11}$

2 $\mathrm{Corr}(V_j, \mathbf{Y}) = \mathbf{b}_j^\top \mathbf{R}_{22}$

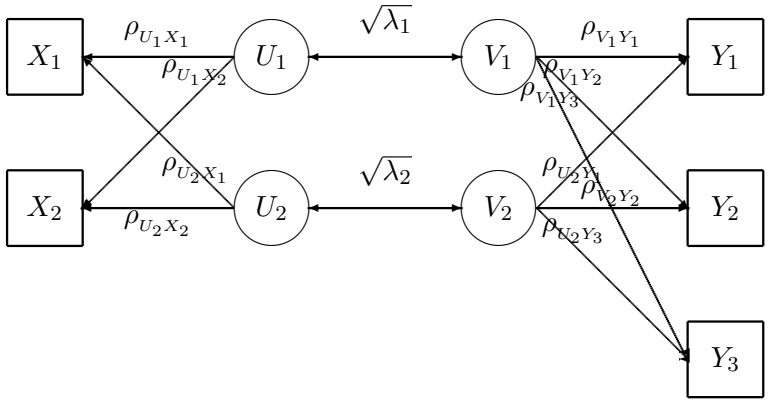


圖 9.1: 典型相關分析路徑圖

每個變數的典型負荷量予以平方，就可獲得每一個原始變數的變異被其典型變量解釋的程度。各變數的典型負荷量平方值的簡單平均數就是典型變量所能解釋之比例，即典型變量自我解釋的能力，可計算如下：

$$\rho_{U_i}^2 = \frac{\sum_{j=1}^{p} \rho_{U_iX_j}^2}{p} \tag{9-11}$$

及

$$\rho_{V_i}^2 = \frac{\sum_{j=1}^{q} \rho_{V_i Y_j}^2}{q} \tag{9-12}$$

由（9-11），可計算前k個典型變量（$U_1, U_2, \ldots U_k$）所解釋$\mathbf{X}$這一組變數之比例爲

$$\rho_{\mathbf{X}|U_1,U_2,\ldots U_k}^2 = \sum_{i=1}^{k} \frac{\sum_{j=1}^{p} \rho_{U_i X_j}^2}{p} \tag{9-13}$$

同理，由（9-12），前k個典型變量（$V_1, V_2, \ldots V_k$）所解釋$\mathbf{Y}$這一組變數之比例爲

$$\rho_{\mathbf{Y}|V_1,V_2,\ldots V_k}^2 = \sum_{i=1}^{k} \frac{\sum_{j=1}^{q} \rho_{V_i Y_j}^2}{q} \tag{9-14}$$

若想了解兩組變數間之整體相關程度，則必須藉由重疊指標的（redundancy index）計算。重疊指標乃指一組變數的變異數中，可被另一組變數的典型變量所能解釋之比例。由於V_i並無法直接對變數$\mathbf{X}$作解釋，因此，必須透過相對應之典型變量U_i才可與變數$\mathbf{X}$連接。$\mathbf{X}$變數組能被$\mathbf{Y}$變數組之典型變量V_i所能解釋之比例可計算如下：

$$\rho_{\mathbf{X}|V_i}^2 = \lambda_i \times \frac{\sum_{j=1}^{p} \rho_{U_i X_j}^2}{p} \tag{9-15}$$

同理，$\mathbf{Y}$變數組能被$\mathbf{X}$變數組之典型變量U_i所能解釋之比例爲

$$\rho_{\mathbf{Y}|U_i}^2 = \lambda_i \times \frac{\sum_{j=1}^{q} \rho_{V_i X_j}^2}{q} \tag{9-16}$$

一般而言，重疊指標若未達5%，則此線性組合之解釋能力即不予慮。由上述說明也可發現，典型相關係數只表達同一對典型變量間的關聯，而重疊指數表達的才是典型變量與另一組變數間的關係。

例 9-1. 已知成人學習動機，包括職業進展、社交關係、外界期望、求知樂趣和自我實現五個控制變項（X_1, X_2, $\ldots X_5$）與學習行爲的教學導向學習和意義導向學習、計劃學習、偶發學習（Y_1, Y_2, Y_3, Y_4）之相關係

數矩陣爲

$$\begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{array} \begin{array}{c} \begin{array}{ccccccccc} X_1 & X_2 & X_3 & X_4 & X_5 & Y_1 & Y_2 & Y_3 & Y_4 \end{array} \\ \left(\begin{array}{ccccccccc} 1.000 & 0.220 & 0.219 & 0.326 & 0.324 & 0.378 & 0.316 & 0.349 & 0.354 \\ 0.220 & 1.000 & 0.495 & 0.502 & 0.489 & 0.387 & 0.437 & 0.394 & 0.386 \\ 0.219 & 0.495 & 1.000 & 0.417 & 0.441 & 0.320 & 0.418 & 0.312 & 0.335 \\ 0.326 & 0.502 & 0.417 & 1.000 & 0.670 & 0.554 & 0.457 & 0.520 & 0.546 \\ 0.324 & 0.489 & 0.441 & 0.670 & 1.000 & 0.500 & 0.466 & 0.610 & 0.608 \\ 0.378 & 0.387 & 0.320 & 0.554 & 0.500 & 1.000 & 0.479 & 0.573 & 0.631 \\ 0.316 & 0.437 & 0.418 & 0.457 & 0.466 & 0.479 & 1.000 & 0.484 & 0.588 \\ 0.349 & 0.394 & 0.312 & 0.520 & 0.610 & 0.573 & 0.484 & 1.000 & 0.580 \\ 0.354 & 0.386 & 0.335 & 0.546 & 0.608 & 0.631 & 0.588 & 0.58 & 1.000 \end{array} \right) \end{array}$$

試進行成人學習動機與學習行爲之整體相關分析。

經計算可得$\mathbf{R}_{xx}^{-\frac{1}{2}}$與$\mathbf{R}_{yy}^{-\frac{1}{2}}$分別爲

$$\mathbf{R}_{xx}^{-\frac{1}{2}} = \begin{bmatrix} 1.0604 & -0.0293 & -0.0461 & -0.1072 & -0.1035 \\ -0.0293 & 1.2141 & -0.2233 & -0.1876 & -0.1574 \\ -0.0461 & -0.2233 & 1.1653 & -0.1009 & -0.1413 \\ -0.1072 & -0.1876 & -0.1009 & 1.3386 & -0.4031 \\ -0.1035 & -0.1574 & -0.1413 & -0.4031 & 1.3397 \end{bmatrix}$$

及

$$\mathbf{R}_{yy}^{-\frac{1}{2}} = \begin{bmatrix} 1.3060 & -0.1265 & -0.2489 & -0.3169 \\ -0.1265 & 1.2258 & -0.1556 & -0.284 \\ -0.2489 & -0.1556 & 1.2647 & -0.2327 \\ -0.3169 & -0.2840 & -0.2327 & 1.3862 \end{bmatrix}$$

故得

$$\mathbf{R}_{xx}^{-\frac{1}{2}}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{R}_{xx}^{-\frac{1}{2}} = \begin{bmatrix} 0.083 & 0.0744 & 0.0560 & 0.1192 & 0.1216 \\ 0.0744 & 0.0831 & 0.0718 & 0.1000 & 0.1034 \\ 0.0560 & 0.0718 & 0.0701 & 0.0730 & 0.0674 \\ 0.1192 & 0.1000 & 0.073 & 0.1748 & 0.1739 \\ 0.1216 & 0.1034 & 0.0674 & 0.1739 & 0.2702 \end{bmatrix}$$

由於$\mathbf{R}_{xx}^{-\frac{1}{2}}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{R}_{xx}^{-\frac{1}{2}}$之特徵根與特徵量分別爲

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \end{bmatrix} = \begin{bmatrix} 0.5675 \\ 0.0797 \\ 0.0323 \\ 0.0017 \\ 0.0000 \end{bmatrix}$$

以及

$$[\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4, \mathbf{e}_5] = \begin{bmatrix} -0.3708 & -0.1634 & -0.2873 & -0.3107 & 0.8104 \\ -0.3393 & -0.4193 & 0.3035 & -0.6801 & -0.393 \\ -0.2543 & -0.5610 & 0.4901 & 0.5925 & 0.1714 \\ -0.5277 & -0.1598 & -0.6713 & 0.293 & -0.3994 \\ -0.6358 & 0.6762 & 0.3667 & 0.0640 & 0.0000 \end{bmatrix}$$

接下來我們可由$\mathbf{a}_i = \mathbf{R}_{11}^{-\frac{1}{2}}\mathbf{e}_i$求得$\mathbf{X}$這組典型變量$\mathbf{U}$的權重如下：

$$\begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \mathbf{a}_3 & \mathbf{a}_4 & \mathbf{a}_5 \end{bmatrix} = \begin{bmatrix} -0.2492 & -0.1879 & -0.3022 & -0.3749 & 0.9058 \\ -0.1453 & -0.4554 & 0.3357 & -1.0139 & -0.4642 \\ -0.0603 & -0.632 & 0.5325 & 0.818 & 0.2904 \\ -0.3210 & -0.3338 & -1.1220 & 0.4674 & -0.5651 \\ -0.5113 & 1.1325 & 0.6746 & 0.0231 & 0.1148 \end{bmatrix}$$

又由(9-9)及$\mathbf{b}_i = \mathbf{R}_{22}^{-\frac{1}{2}}\mathbf{f}_i$，可求得$\mathbf{Y}$這組典型變量$\mathbf{V}$爲

$$\begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \mathbf{b}_3 & \mathbf{b}_4 \end{bmatrix} = \begin{bmatrix} 0.2468 & 0.6411 & -1.1674 & -0.2218 \\ 0.2152 & 1.0018 & 0.7573 & 0.0125 \\ 0.4057 & -0.6546 & 0.3843 & -0.9995 \\ 0.3461 & -0.7991 & 0.0603 & 1.1810 \end{bmatrix}$$

最後，我們可得$\mathbf{U}$與$\mathbf{X}$的相關係數矩陣與$\mathbf{V}$與$\mathbf{Y}$的相關係數矩陣分別爲

$$\begin{bmatrix} \mathbf{a}_1^\top\mathbf{R}_{xx} & \mathbf{a}_2^\top\mathbf{R}_{xx} & \mathbf{a}_3^\top\mathbf{R}_{xx} & \mathbf{a}_4^\top\mathbf{R}_{xx} \end{bmatrix} = \begin{bmatrix} -0.5647 & -0.1684 & -0.2589 & -0.2589 \\ -0.6411 & -0.4234 & 0.2994 & -0.4455 \\ -0.5462 & -0.5383 & 0.4621 & 0.4392 \\ -0.8429 & -0.1284 & -0.3780 & 0.1929 \\ -0.9048 & 0.3466 & 0.2239 & 0.0798 \end{bmatrix}$$

以及

$$\left[\begin{array}{cccc} \mathbf{b}_1^\top \mathbf{R}_{yy} & \mathbf{b}_2^\top \mathbf{R}_{yy} & \mathbf{b}_3^\top \mathbf{R}_{yy} & \mathbf{b}_4^\top \mathbf{R}_{yy} \end{array}\right] = \left[\begin{array}{cccc} 0.8007 & 0.2416 & -0.5464 & -0.0433 \\ 0.7333 & 0.5221 & 0.4195 & 0.1170 \\ 0.8520 & -0.2659 & 0.1168 & -0.4356 \\ 0.8637 & -0.1852 & -0.0082 & 0.4687 \end{array}\right]$$

9.2 典型變量個數之選取

典型相關分析的主要目的爲探討兩組變數組間的整體相關形式。析與主成分分析、因素分析相似，皆可作爲降低維度的方法。然而，到底需要幾個典型變量才足夠，這沒有一定的標準答案。典型變量個數之選取有以下幾種方式：

1 依照文獻決定有多少組典型變量。

2 依照統計結果分析決定有多少組典型變量。

3 以統計檢定的方式決定有多少組典型變量，通常採用Bartlett於1942年所提出的概度比檢定，其統計假設及檢定統計量分別爲

$$\begin{cases} H_0 : \lambda_{k+1} = \lambda_{k+2} = \cdots = \lambda_p = 0 \\ H_1 : \lambda_i \text{不全爲} 0, i = k+1, k+2, \ldots p \end{cases}$$

$$V = -\left(n - 1 - \frac{1}{2}(p+q+1)\right) \ln \prod_{i=k+1}^{p} (1-\lambda_i) \sim \chi^2(pq) \tag{9-17}$$

所以，當$k = 0$時，即爲檢定所有特徵值是否均相等。若不棄卻虛無假設，則表示所有特徵值均爲0，因此不需作典型相關分析。此一檢定爲連續檢定$k = 0, 1, 2, \ldots, p$直到棄卻虛無假設爲止。

9.3 範例

試以下列資料進行典型相關分析

X_1	50	42	61	35	50	39	53	28	23	22	70	76	43	40	38
X_2	70	89	47	48	58	42	71	49	42	59	91	62	53	59	61
Y_1	86	77	86	63	83	62	85	70	63	72	84	85	78	71	79
Y_2	89	78	84	67	86	73	91	69	50	69	87	89	88	67	75

由於SPSS並無典型相關分析之操作介面，所以必須藉由執行程式語法之功能才可執行。SPSS操作步驟如下：

Step 1. SPSS功能列⇒檔案⇒開新檔案⇒語法

Step 2. 典型相關分析語法如下，輸入後選擇執行全部即可。

```
Include file 'C:\Program files\SPSS\canonical correlation.sps'.
Cancorr
set1=x1,x2/
set2=y1,y2/.
```

SPSS報表輸出$X1$與$X2$的相關係數矩陣及共變異數矩陣如下：

Correlations for Set-1

	X1	X2
X1	1.0000	0.4569
X2	0.4569	1.0000

Correlations for Set-2

	Y1	Y2
Y1	1.0000	0.8623
Y2	0.8623	1.0000

Correlations Between Set-1 and Set-2

	Y1	Y2
X1	0.7611	0.7876
X2	0.5661	0.5013

由相關係數矩陣可計算得兩典型變量之相關係數$\sqrt{\lambda_1}$及$\sqrt{\lambda_2}$分別爲0.829及0.172。

Canonical Correlations

1	0.829
2	0.172

在特徵值檢定方面，首先檢定

$$\begin{cases} H_0 : \lambda_1 = \lambda_2 = 0 \\ H_1 : \lambda_i \text{不全爲} 0, i = 1, 2 \end{cases}$$

由於P-Value=0.08<0.05，因此經Bartlett檢定結果棄卻虛無假設；即至少一組典型相關係數不爲0。在第二次檢定

$$\begin{cases} H_0 : \lambda_2 = 0 \\ H_1 : \lambda_2 \neq 0 \end{cases}$$

由於P-Value=0.557>0.05，因此不棄卻虛無假設；即表示第二組典型相關係數爲0。因此，此範例僅需一組典型變量。

Test that remaining correlations are zero				
	Wilk's	Chi-SQ	DF	Sig.
1	0.303	13.741	4.000	0.008
2	0.970	0.346	1.000	0.557

根據上述λ_1及λ_2可求得各組典型權重$\mathbf{a}_1$、$\mathbf{a}_2$、$\mathbf{b}_1$及$\mathbf{b}_2$如下：

Standardized Canonical Coefficients for Set-1		
	$\mathbf{a}_1$	$\mathbf{a}_2$
X1	−0.839	−0.748
X2	−0.282	1.088

Standardized Canonical Coefficients for Set-2		
	$\mathbf{b}_1$	$\mathbf{b}_2$
Y1	−0.501	1.910
Y2	−0.536	−1.900

此外，**X**組變項之典型變量與各變項之相關係數，即典型負荷量如下：

Canonical Loadings for Set-1		
	U_1	U_2
X1	−0.968	−0.251
X2	−0.665	0.746

而**Y**組變項之典型變量與各變項之典型負荷量爲

Canonical Loadings for Set-2		
	V_1	V_2
Y1	−0.963	0.271
Y2	−0.967	−0.254

因此，由下表可知，**X**組變項之第一組典型變量（U_1）能解釋**X**組變項之變異比例爲0.69。而**Y**組變項之第一組典型變量（V_1）透過U_1能解釋**X**組變項之變異比例爲0.475。

Redundancy Analysis	
Proportion of Variance of Set-1 Explained by Its Own Can. Var.	
	Prop Var
CV1-1	0.690
CV1-2	0.310

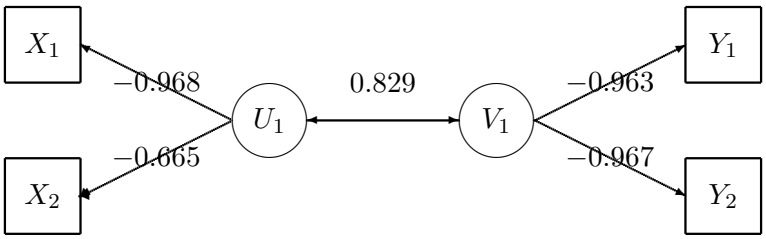
Redundancy Analysis	
Proportion of Variance of Set-1 Explained by Opposite Can. Var.	
	Prop Var
CV2-1	0.475
CV2-2	0.009

同理，**Y**組變項之第一組典型變量（V_1）能解釋**Y**組變項之變異比例爲0.931。而**X**組變項之第一組典型變量（U_1）透過V_1能解釋**Y**組變項之變異比例爲0.641。

Redundancy Analysis	
Proportion of Variance of Set-2 Explained by Its Own Can. Var.	
	Prop Var
CV2-1	0.931
CV2-2	0.069

Redundancy Analysis	
Proportion of Variance of Set-2 Explained by Opposite Can. Var.	
	Prop Var
CV1-1	0.641
CV1-2	0.002

最後，我們可繪製典型分析路徑圖如下：



第 10 章

區別分析

區別分析（Discriminant Analysis，或稱鑑別分析）是一種相依方法，其準則變數爲事先訂定的類別或組別。區別分析主要是建一組區別函數或區別規則，使得資根據它們分類與原先的分類最爲接近。舉例來說，銀行在借錢給顧客時，通常都會根據已存在顧客的基本資料（分爲具信用之顧客與不具信用之顧客兩種），包含學歷、收入、借貸記錄...，做一區別分析；並且當有新的顧客來時，我們可將這新的顧客的資料與這些已存在的資料做一比較，觀看是否應貸款給這位新顧客。

10.1 兩群體區別分析

設$f_i(\mathbf{x}),\ i=1,2$分別爲群體$\pi_1,\ \pi_2$之機率密函數，其中$\mathbf{X}$爲$p\times 1$的可觀測隨機向量。因此，當樣本屬π_1而歸類給π_2之條件機率爲

$$P(2|1)=P(\mathbf{X}\in\mathbf{R}_2|\pi_1)=\int_{\mathbf{R}_2=\Omega-\mathbf{R}_1}f(\mathbf{x})d\mathbf{x}$$

同理，當樣本屬π_2而歸類給π_1之條件機率爲

$$P(1|2)=P(\mathbf{X}\in\mathbf{R}_1|\pi_2)=\int_{\mathbf{R}_1=\Omega-\mathbf{R}_2}f(\mathbf{x})d\mathbf{x}$$

若假設$p_i,\ i=1,2$分爲群體π_i之事前機率（prior probability）且$p_1+p_2=1$，則將一樣本歸類正確與否有以下四種情況；

1 將樣本正確歸類給π_1，以機率可表示爲

$$P(\text{將樣本正確歸類給}\pi_1)=P(\pi_1)\times P(\mathbf{X}\in\mathbf{R}_1|\pi_1)$$

2 將樣本錯誤歸類給π_1，以機率可表示爲

$$P(\text{將樣本錯誤歸類給}\pi_1) = P(\pi_2) \times P(\mathbf{X} \in \mathbf{R}_1|\pi_2)$$

3 將樣本正確歸類給π_2，以機率可表示爲

$$P(\text{將樣本正確歸類給}\pi_2) = P(\pi_2) \times P(\mathbf{X} \in \mathbf{R}_2|\pi_2)$$

4 將樣本錯誤歸類給π_2，以機率可表示爲

$$P(\text{將樣本正確歸類給}\pi_2) = P(\pi_1) \times P(\mathbf{X} \in \mathbf{R}_2|\pi_1)$$

可求得總誤歸率TPM（total probability of misclassification）爲

$$\text{TPM} = p_1 P(2|1) + p_2 P(1|2) \tag{10-1}$$

由於錯誤的歸類可能造成決策者成本的損失，在此定義誤歸成本如下：

		眞實狀況	
		π_1	π_2
決策	π_1	0	$c(1\|2)$
	π_2	$c(2\|1)$	0

由期望值之定義，可求得期望誤歸成本（expected cost of misclassification）ECM爲

$$\text{ECM} = p_1 P(2|1) c(2|1) + p_2 P(1|2) c(1|2) \tag{10-2}$$

一般區別分析方法有以下四種，期望誤歸成本法、總誤歸率法、貝氏事後機率法及一般平方距離法以，茲分別介紹如下：

1 期望誤歸成本法：爲找到最佳的決策使得ECM爲最小，其最佳區別法則爲：

$$\begin{cases} \mathbf{R}_1 : \dfrac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \dfrac{c(1|2)}{c(2|1)} \dfrac{p_2}{p_1} \\ \mathbf{R}_2 : \dfrac{f_2(\mathbf{x})}{f_1(\mathbf{x})} \geq \dfrac{c(2|1)}{c(1|2)} \dfrac{p_1}{p_2} \end{cases} \tag{10-3}$$

2 總誤歸率法：爲找到最佳的決策使得TPM爲最小，由ECM之定義可知，當誤判成本$c(1|2) = c(2|1)$時，TPM法則即爲ECM法則之特例。因此，可求得其最佳區別法則爲：

$$\begin{cases} \mathbf{R}_1 : \dfrac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \dfrac{p_2}{p_1} \\ \\ \mathbf{R}_2 : \dfrac{f_2(\mathbf{x})}{f_1(\mathbf{x})} \geq \dfrac{p_1}{p_2} \end{cases} \tag{10-4}$$

3 貝氏事後機率法：以觀測值之事後機率作爲歸類之標準。由貝式法則可知

$$\begin{aligned} P(\mathbf{X} \in \pi_1 | \mathbf{X} = \mathbf{x}) &= \frac{P(\mathbf{X} \in \pi_1 \cap \mathbf{X} = \mathbf{x})}{P(\mathbf{X} = \mathbf{x})} \\ &= \frac{P(X \in \pi_1 \cap X = \mathbf{x})}{P(X \in \pi_1 \cap X = \mathbf{x}) + P(X \in \pi_2 \cap X = \mathbf{x})} \\ &= \frac{p_1 P(X = \mathbf{x}|\pi_1)}{p_1 P(X = \mathbf{x}|\pi_1) + p_1 P(X = \mathbf{x}|\pi_2)} \\ &= \frac{p_1 f_1(\mathbf{x})}{p_1 f_1(\mathbf{x}) + p_2 f_2(\mathbf{x})} \end{aligned}$$

以及

$$P(\mathbf{X} \in \pi_2 | \mathbf{X} = \mathbf{x}) = \frac{p_1 f_1(\mathbf{x})}{p_1 f_1(\mathbf{x}) + p_2 f_2(\mathbf{x})}$$

因此，區別法則爲：

$$\begin{cases} \mathbf{R}_1 : p_1 f_1(\mathbf{x}) \geq p_2 f_2(\mathbf{x}) \\ \\ \mathbf{R}_2 : p_2 f_2(\mathbf{x}) < p_1 f_1(\mathbf{x}) \end{cases} \tag{10-5}$$

4 一般平方距離法：以統計距離遠近作爲歸類之標準。觀測值$\mathbf{x}$與群體π_i之統計距離定義爲

$$D_i(\mathbf{x}|i) = (\mathbf{x} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i), i = 1, 2 \tag{10-6}$$

統計距離越大，表示觀測值$\mathbf{x}$與第π_i之群中點越遠，因此，區別法則爲：

$$\begin{cases} \mathbf{R}_1 : D_1(\mathbf{x}|1) \leq D_2(\mathbf{x}|2) \\ \\ \mathbf{R}_2 : D_1(\mathbf{x}|1) > D_2(\mathbf{x}|2) \end{cases} \tag{10-7}$$

由以上介紹可發現，在不考慮誤歸成本情況下（即$c(1|2) = c(2|1)$），根據ECM、TPM及貝氏事後機率法則所得之區別法則爲一致的。

10.1.1 資料來自常態母體，且共變異數矩陣相等

而當資料爲來自常態母體，且共變異數矩陣相等（$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$）時，由多變量常態分配之機率密度函數可知

$$\begin{aligned}\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} &= \frac{\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right]}{\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right]} \\ &= \exp\left[\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right]\end{aligned}$$

由（10-3），因爲

$$\begin{aligned}&\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) \\ =\ &(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}\mathbf{x}-\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\end{aligned}$$

故區別法則爲：

$$\begin{cases}\mathbf{R}_1:(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}\mathbf{x}-\frac{1}{2}(\boldsymbol{\mu}_1-\mu_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\geq\ln\left[\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}\right] \\ \mathbf{R}_2:(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}\mathbf{x}-\frac{1}{2}(\boldsymbol{\mu}_1-\mu_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)<\ln\left[\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}\right]\end{cases} \tag{10-8}$$

在實務上，由於母體均值向量及共變異數矩陣未知，通常以

$$\overline{\mathbf{x}}_i \text{ 和 } \mathbf{S}_{\text{pooled}}=\frac{(n_1-1)\mathbf{S}_1+(n_2-1)\mathbf{S}_2}{n_1+n_2-2}$$

作爲群體均值向量及共同母體共變異數矩陣之估計值。

此外，在不考慮誤歸成本及事前機率情況下（即$c(1|2)=c(2|1)$，$p_1=p_2$）

$$\begin{aligned}\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} &= \frac{\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_1|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right]}{\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_2|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right]} \\ &= \frac{\exp\left[-D_1(\mathbf{x}|1)\right]}{\exp\left[-D_2(\mathbf{x}|2)\right]}\end{aligned}$$

根據ECM、TPM法則所得之區別法則爲

$$\begin{cases}\mathbf{R}_1:D_1(\mathbf{x}|1)<D_2(\mathbf{x}|2) \\ \mathbf{R}_2:D_1(\mathbf{x}|1)>D_2(\mathbf{x}|2)\end{cases} \tag{10-9}$$

綜合以上討論，讀者可發現在此種情況下，不論以機率密度、事後機率或距離的觀點所得到的區別法則皆爲一致的。

10.1.2 資料來自常態母體，但共變異數矩陣不相等

而當資料爲來自常態母體，但共變異數矩陣不相等（$\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$）時，由多變量常態分配之機率密度函數可知

$$\begin{aligned}\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} &= \frac{\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_1|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right]}{\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_2|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right]} \\ &= \frac{|\boldsymbol{\Sigma}_2|^{1/2}}{|\boldsymbol{\Sigma}_1|^{1/2}}\exp\left[\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right]\end{aligned}$$

又

$$\begin{aligned}&\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) \\ =\ & -\frac{1}{2}\mathbf{x}^\top\left(\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\Sigma}_2^{-1}\right)\mathbf{x}-\left(\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1}\right)^\top\mathbf{x}\end{aligned}$$

根據ECM法則所得之區別法則爲

$$\begin{cases}\mathbf{R}_1: -\frac{1}{2}\mathbf{x}^\top\left(\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\Sigma}_2^{-1}\right)\mathbf{x}-\left(\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1}\right)^\top\mathbf{x}-k \geq \ln\left[\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}\right] \\ \mathbf{R}_2: -\frac{1}{2}\mathbf{x}^\top\left(\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\Sigma}_2^{-1}\right)\mathbf{x}-\left(\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1}\right)^\top\mathbf{x}-k < \ln\left[\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}\right]\end{cases} \tag{10-10}$$

其中

$$k=\frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}+\frac{1}{2}\left(\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1-\boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\right)$$

10.1.3 共變異數矩陣同質性檢定

SPSS中進行區別分析時，可以Box's M來檢定多個母體之共變異數矩陣是否相等，其統計假設及檢定統計量分別爲

$$\begin{cases}H_0: \boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\cdots=\boldsymbol{\Sigma}_g \\ H_1: \boldsymbol{\Sigma}_i\text{不全相等}\end{cases}$$

和

$$M = \gamma \left[(n-g) \log |\mathbf{S}| - \sum_{i=1}^{g} (n_i - 1) \log |\mathbf{S}_i| \right] \sim F(v_1, v_2) \tag{10-11}$$

其中

$$\begin{aligned}
f_1 &= \frac{p(p+1)(g-1)}{2} \\
f_2 &= \frac{f1+2}{\left|\tau - (1-\rho)^2\right|} \\
\rho &= 1 - \frac{2p^2+3p-1}{6(p+1)(g-1)} \left(\sum_{i=1}^{g} \frac{1}{n_i - 1} - \frac{1}{n-g} \right) \\
\tau &= \frac{(p-1)(p+1)}{6(g-1)} \left(\sum_{i=1}^{g} \frac{1}{(n_i-1)^2} - \frac{1}{(n-g)^2} \right) \\
\gamma &= \frac{\rho - f_1/f_2}{f_1}
\end{aligned}$$

當$M > F_\alpha(v_1, v_2)$時，則棄卻虛無假設。

10.2 多群體區別分析

設$f_i(\mathbf{x})$分別爲群體π_1, π_2, ..., π_g之機率密函數，其中$\mathbf{X}$爲$p \times 1$的可觀測隨機向量。可求得(x)眞正隸屬第k個群體，而被誤歸爲其他群體之ECM及TPM分別爲

$$\mathrm{ECM}(k) = \sum_{i=1}^{g} p_k P(i|k) c(i|k) \tag{10-12}$$

$$\mathrm{TPM}(k) = \sum_{i=1}^{g} p_k P(i|k) \tag{10-13}$$

按上節所介紹之五種區別方法所得的區別法則分別如下：

1 期望誤歸成本法：$\mathbf{x}$隸屬第k個群體，若

$$\sum_{i=1}^{g} p_k P(i|k) c(i|k) = \min_j \sum_{i=1; i \neq k}^{g} p_k P(i|j) c(i|j) \tag{10-14}$$

2 總誤歸率法：$\mathbf{x}$隸屬第k個群體，若

$$\sum_{i=1}^{g} p_k P(i|k) = \min_j \sum_{i=1;i\neq k}^{g} p_k P(i|j) \tag{10-15}$$

3 貝氏事後機率法：$\mathbf{x}$隸屬第k個群體，若

$$P\left(\mathbf{X} \in \pi_k | \mathbf{X} = \mathbf{x}\right) = \max_j \frac{p_j f_j\left(\mathbf{x}\right)}{\sum_{i=1}^{g} p_i f_i\left(\mathbf{x}\right)}$$

4 一般平方距離法：$\mathbf{x}$隸屬第k個群體，若

$$D_k(\mathbf{x}|k) = \min_j D_j(\mathbf{x}|j) \tag{10-16}$$

而當資料爲來自常態母體且及不考慮誤歸成本時，由於

$$\ln p_i f_i(\mathbf{x}) = \ln p_i - \frac{p}{2}\ln\left(2\pi\right) - \frac{1}{2}\left|\mathbf{\Sigma}_i\right| - \frac{1}{2}\left(\mathbf{x}-\mu_i\right)^\top \mathbf{\Sigma}_i^{-1}\left(\mathbf{x}-\mu_i\right) \tag{10-17}$$

上式中$-p/2\ln\left(2\pi\right)$對比較$\ln p_i f_i(\mathbf{x})$大小並無影響；因此，在母體共同變異數矩陣不相等時，區別法則爲：$\mathbf{x}$隸屬第k個群體，若

$$d_k^Q\left(\mathbf{x}\right) = \max_j d_j^Q\left(\mathbf{x}\right) \tag{10-18}$$

其中

$$d_i^Q\left(\mathbf{x}\right) = \ln p_i - \frac{1}{2}\left|\mathbf{\Sigma}_i\right| - \frac{1}{2}\left(\mathbf{x}-\mu_i\right)^\top \mathbf{\Sigma}_i^{-1}\left(\mathbf{x}-\mu_i\right) \tag{10-19}$$

由於$d^Q\left(\mathbf{x}\right)$中爲$\mathbf{x}$之二次式，故以二次區別函數稱之。將觀測值$\mathbf{x}$帶入所求得之值稱爲二次區別分數。

上式中$d^Q\left(\mathbf{x}\right)$可改寫爲

$$\begin{aligned} d^Q\left(\mathbf{x}\right) &= \ln p_i - \frac{1}{2}\left|\mathbf{\Sigma}_i\right| - \frac{1}{2}\left(\mathbf{x}-\mu_i\right)^\top \mathbf{\Sigma}_i^{-1}\left(\mathbf{x}-\mu_i\right) \\ &= \ln p_i - \frac{1}{2}\left|\mathbf{\Sigma}_i\right| - \frac{1}{2}\mathbf{x}^\top\mathbf{\Sigma}_i^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i^\top\Sigma_i^{-1}\boldsymbol{\mu}_i + \boldsymbol{\mu}_i^\top\mathbf{\Sigma}_i^{-1}\mathbf{x} \end{aligned} \tag{10-20}$$

當群體共變異數矩陣相等（$\mathbf{\Sigma}_1 = \cdots = \mathbf{\Sigma}_g = \mathbf{\Sigma}$），$-1/2\left|\mathbf{\Sigma}\right|$及$-\frac{1}{2}\mathbf{x}^\top\mathbf{\Sigma}^{-1}\mathbf{x}$對比較$\ln p_i f_i(\mathbf{x})$大小亦無影響；所以，當群體共同變異數矩陣相等時，區別法則爲：$\mathbf{x}$隸屬第k個群體，若

$$d_k\left(\mathbf{x}\right) = \max_j d_j\left(\mathbf{x}\right) \tag{10-21}$$

其中

$$d_i\left(\mathbf{x}\right) = \ln p_i - \frac{1}{2}\boldsymbol{\mu}_i^\top\Sigma^{-1}\boldsymbol{\mu}_i + \boldsymbol{\mu}_i^\top\mathbf{\Sigma}^{-1}\mathbf{x} \tag{10-22}$$

由於$d(\mathbf{x})$中爲$\mathbf{x}$之線性組合，故以線性區別函數稱之。將觀測值$\mathbf{x}$帶入所求得之值稱爲線性區別分數。 實務上，群體共變異數矩陣未知，在此以

$$\mathbf{S}_{\text{pooled}}=\frac{\sum_{i=1}^{g}(n_i-1)\,\mathbf{S}_i}{n_1+n_2+\cdots+n_g-g}$$

作爲共同母體共變異數矩陣之估計值。

10.3 典型區別分析

典型區別分析（Canonical Discriminant Analysis）爲Fisher於1936年所提出，假設各群體共變異數矩陣相等條件下，尋找區別變數的線性組合之最佳權重，使組間變異對組內變異比值爲最大。換句話說，即尋找使群體間差異最大的線性組合，如此，同一群的資料能盡量集中，而不同群之間的資料能盡量分開。令$Y=\mathbf{a}^{\top}\mathbf{X}$，由於$\mathrm{E}Y=\mathbf{a}^{\top}\boldsymbol{\mu}$及$\mathrm{Var}(Y)=\mathbf{a}^{\top}\boldsymbol{\Sigma}\mathbf{a}$，故可得

$$\frac{\mathbf{Y}\text{的組間變異量}}{\mathbf{Y}\text{的變異量}}=\frac{\mathbf{a}^{\top}\mathbf{B}\mathbf{a}}{\mathbf{a}^{\top}\boldsymbol{\Sigma}\mathbf{a}}=\frac{\mathbf{a}^{\top}\left(\sum_{i=1}^{g}(\mu_i-\overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_i-\overline{\boldsymbol{\mu}})^{\top}\right)\mathbf{a}}{\mathbf{a}^{\top}\boldsymbol{\Sigma}\mathbf{a}}$$

其中

$$\overline{\boldsymbol{\mu}}=\frac{1}{g}\sum_{i=1}^{g}\boldsymbol{\mu}_i$$

由於群體變異數矩陣未知，因此以樣本共變異數矩陣$\mathbf{S}_{\text{pooled}}$作爲$\boldsymbol{\Sigma}$估計值。所以，目標函數即爲

$$\max_{\mathbf{a}}\quad\frac{\mathbf{a}^{\top}\mathbf{B}\mathbf{a}}{\mathbf{a}^{\top}\mathbf{S}_{\text{polled}}\mathbf{a}}\tag{10-23}$$

因爲

$$\begin{aligned}(n_1+n_2+\cdots+n_g-g)\mathbf{S}_{\text{pooled}}&=\sum_{i=1}^{g}(n_i-1)\,\mathbf{S}_i\\&=\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{x}_{ij}-\overline{\mathbf{x}}_i)(\mathbf{x}_{ij}-\overline{\mathbf{x}}_i)^{\top}\\&=\mathbf{W}\end{aligned}$$

所以，可發現求解（10-23）與求解下列問題所得到的$\mathbf{a}$是一致的

$$\max\quad\frac{\mathbf{Y}\text{的組間變異量}}{\mathbf{Y}\text{的組內變異量}}=\frac{\mathbf{a}^{\top}\mathbf{B}\mathbf{a}}{\mathbf{a}^{\top}\mathbf{W}\mathbf{a}}\tag{10-24}$$

求解上式可得

$$\max \quad \frac{\mathbf{a}_1^\top \mathbf{B}\mathbf{a}_1}{\mathbf{a}_1^\top \mathbf{W}\mathbf{a}_1} = \lambda_1$$

其中，λ_1爲$\mathbf{W}^{-1}\mathbf{B}$之最大特徵值及$\mathbf{a}_1$爲$\lambda_1$所對應之特徵向量。同理，$\lambda_i$爲$\mathbf{W}^{-1}\mathbf{B}$ 之第i大特徵值及$\mathbf{a}_i$爲λ_i所對應之特徵向量，$i = 1, 2, \ldots, s$，s爲$\mathbf{W}^{-1}\mathbf{B}$正特徵值數目且$s \leq \min\{g-1, p\}$。

費雪典型區別分析所得到的每一個典型區別函數，就是經由區別變數轉換得到區別空間上的一個維度。在此一維度上可以解釋原始變數類組間變異的最大部分，特徵值越大，表示對應之函數的區別效力也越大。因此，對應第一維度的區別函數（$\mathbf{a}_1^\top \mathbf{X}$）稱爲第一典型區別函數，他在解釋原始變數類組間變異最具解釋力（即區別效力最大）；然後按照同一原則尋找所建立的第二典型區別函數（$\mathbf{a}_2^\top \mathbf{X}$），他在原始變數類組間變異解釋能力次之，直到推導出所有區別函數。所以後面所推導出的區別函數由於只能解釋很少部分的變異，對於這些實際效果不大的區別函數，則可以被省略。而典型區別變數個數之取法可參考典型相關分析中所介紹之Bartlett球形檢定（9.2），故不再贅述。

典型區別分析是典型相關的特殊情況，即一組變數爲區別變數；另一組變數爲分組變數的情況下所使用的區別分析法。目的在找出區別變數間各種線性組合之典型變數，使其儘可能分組變數之間相關性最高。在此，典型區別分析中所求得之典型相關係數與矩陣$\mathbf{W}^{-1}\mathbf{B}$之特徵值有下列關係

$$\lambda_{\text{dis}} = \frac{\lambda_{\text{can}}}{1 - \lambda_{\text{can}}} \tag{10-25}$$

而當只使用r個典型區別函數時，最佳區別法則爲：$\mathbf{x}$隸屬第k個群體，若

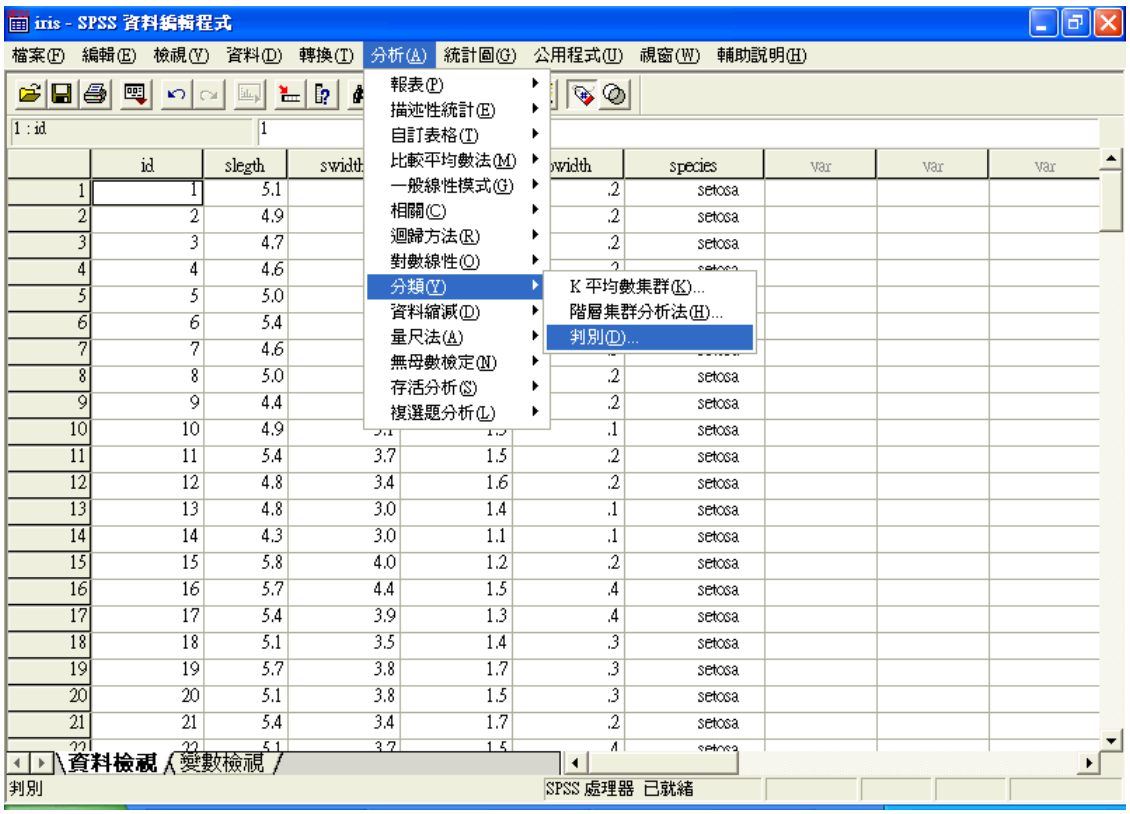
$$\sum_{j=1}^{r} \left[\mathbf{a}_j^\top (\mathbf{x} - \boldsymbol{\mu}_k)\right]^2 = \min_i \sum_{j=1}^{r} \left[\mathbf{a}_j^\top (\mathbf{x} - \boldsymbol{\mu}_i)\right]^2 \tag{10-26}$$

最後必須說明，費雪典型區別分析雖然是建立在等共變異數矩陣的假設下，理論上當Box's M檢定去卻虛無假設時即無法使用。然而當各組的共變異數矩陣不相等時，費雪典型區別分析則不失爲一個次佳的檢定方法。實務上，我們還是可以費雪典型區別分析作爲預設的方法。

10.4 範例

試以Iris資料檔中四個分類變數－花萼長、花萼寬、花瓣長與花瓣寬來當作區別變數；花種爲分類變數進行區別分析。

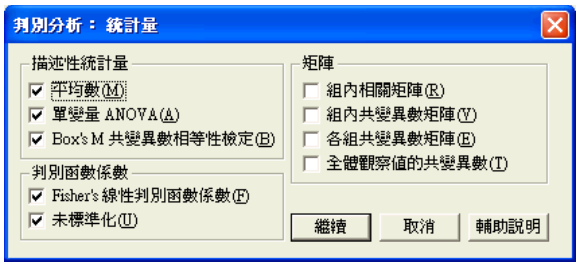
Step 1. SPSS功能列⇒分析⇒分類⇒判別



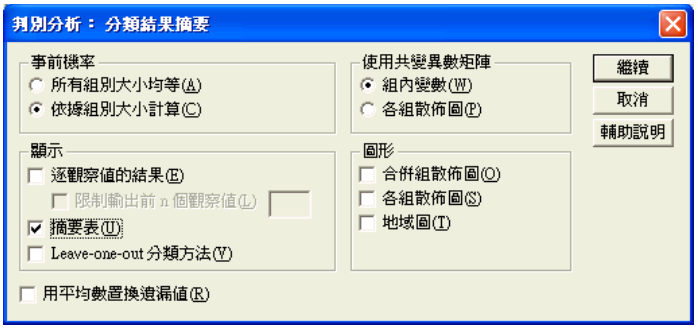
Step 2. 選擇自變數及分組變數



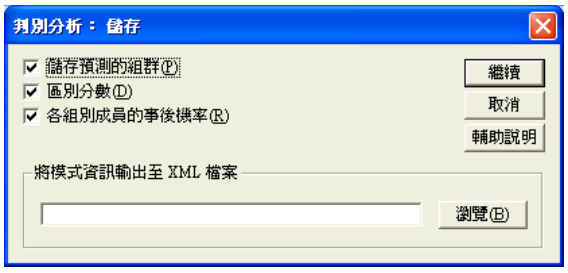
Step 3. 在『統計量』方塊中，選取列出相關係數之『平均數』、『單變量』、『Box's M共變異相等性檢定』、『Fisher 線性區別函數』及『未標準化』。



Step 4. 在『分類』方塊中，選取『依據組別大小計算』決定事前機率並選取輸出『摘要表』。



Step 5. 在『儲存』方塊中，選取『儲存預測的群組』『區別分數』『各組別成員的事後機率』。



SPSS報表輸出如下：

由下表變異數檢定可發現，各區別變數平均數檢定之P-velue均小於0.05，因此不同品種之鳶尾花其花萼長、花萼寬、花瓣長與花瓣寬也會有所差異。由於各區別變數存在顯著性差異，故可進行區別分析。

各組平均數的相等性檢定

	Wilks' Lambda值	F檢定	分子自由度	分母自由度	顯著性
Sepal.Length	0.381	119.265	2	147	0.000
Sepal.Width	0.599	49.160	2	147	0.000
Petal.Length	0.059	1180.161	2	147	0.000
Petal.Width	0.071	960.007	2	147	0.000

由Box's M共變數相等性檢定可發現，P-velue $= 0.000 < .05$，因此棄卻虛無假設

$$\begin{cases} H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}_3 \\ H_1 : \mathbf{\Sigma}_i \text{不全相等} \end{cases}$$

各母體共變異矩陣不相等。由於該檢定對樣本數高度敏感，因此常會導致棄卻虛無假設，通常只要資料不偏離常態假設，仍然可適用。

檢定結果

Box's M共變數相等性檢定		146.663
F檢定	近似值	7.045
	分子自由度	20
	分母自由度	77566.751
	顯著性	0.000

由於區別變數個數$p = 4$，分組變數$g = 3$，故可求得典型變數個數爲$s = \min\{4, 2\} = 2$。兩組典型變量之相關係數分別爲0.985及0.471。由（10-26），可反推得$\mathbf{W}^{-1}\mathbf{B}$之特徵值分別爲32.192及0.285。第一組典型區別函數解釋32.192%（$= \frac{32.192}{32.192+0.285}$）的變異，但第二個判別函數的特徵值爲只解釋0,879%的變異。

特徵值

函數	特徵值	變異數的%	累積%	典型相關
1	32.192	99.121	99.12	0.985
2	0.285	0.879	100.00	0.471

經Bartlett檢定

$$\begin{cases} H_0 : \lambda_1 = \lambda_2 = 0 \\ H_1 : \lambda_i \text{不全爲} 0, i = 1, 2 \end{cases}$$

由於P-value=0.000<0.05，因此棄卻虛無假設；即至少一組典型相關係數不爲0。在第二次檢定

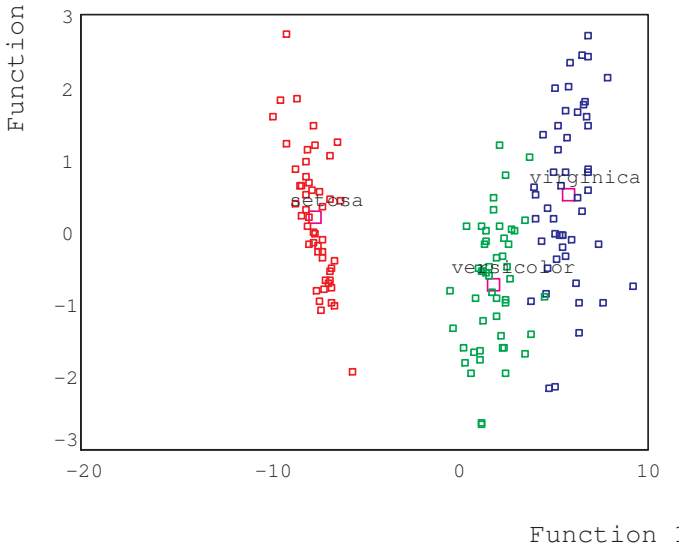
$$\begin{cases} H_0 : \lambda_2 = 0 \\ H_1 : \lambda_2 \neq 0 \end{cases}$$

由於P-value=0.000<0.05，亦棄卻虛無假設；即表示第二組典型相關係數不爲0。因此，此範例需兩組典型變量。

Wilks' Lambda值

函數檢定	Wilks' Lambda值	卡方	自由度	顯著性
1 到 2	0.023	546.115	8	0.000
2	0.778	36.530	3	0.000

根據上述λ_1及λ_2可求得各組典型區別權重$\mathbf{a}_1$、$\mathbf{a}_2$及合併各組之散佈圖如下：



標準化的典型區別函數係數

	函數	
	$\mathbf{a}_1$	$\mathbf{a}_2$
Sepal.Length	−0.427	0.012
Sepal.Width	−0.521	0.735
Petal.Length	0.947	−0.401
Petal.Width	0.575	0.581

此外，典型變量與各區別變項之相關係數爲

結構矩陣

	函數	
	1	2
Petal.Length	0.706	0.168
Sepal.Width	−0.119	0.864
Petal.Width	0.633	0.737
Sepal.Length	0.223	0.311

由上表可發現，第一組典型區別函數與花萼長有高度相關，故可命名爲「花萼長」。而第二組典型區別函數則與其它三個花的特性有高度相關，故可命名爲「其它」。

下表爲各分組觀察值的中心位置，它是利用將典型區別函數代入所有區別變數的平均值所計算出。

各組重心的函數

species	函數 1	函數 2
setosa	−7.608	0.215
versicolor	1.825	−0.728
virginica	5.783	0.513

分類函數係數

	species setosa	species versicolor	species virginica
Sepal.Length	23.544	15.698	12.446
Sepal.Width	23.588	7.073	3.685
Petal.Length	−16.431	5.211	12.767
Petal.Width	−17.398	6.434	21.079
(常數)	−86.308	−72.853	−104.368

Fisher's線性區別函數

由『分類函數係數表』可知，各組之線性區別函數如下：

$$\begin{aligned}\text{setosa} \quad = \quad & -86.308 + 23.544 \times \text{Sepal.Length} + 23.588 \times \text{Sepal.Width} \\ & -16.431 \times \text{Petal.Length} - 17.398 \times \text{Petal.Width}\end{aligned}$$

$$\begin{aligned}\text{versicolor} \quad = \quad & -72.853 + 15.698 \times \text{Sepal.Length} + 7.073 \times \text{Sepal.Width} \\ & +5.211 \times \text{Petal.Length} + 6.434 \times \text{Petal.Width}\end{aligned}$$

$$\begin{aligned}\text{virginica} \quad = \quad & -104.368 + 12.446 \times \text{Sepal.Length} + 3.685 \times \text{Sepal.Width} \\ & +12.767 \times \text{Petal.Length} + 21.079 \times \text{Petal.Width}\end{aligned}$$

從摘要表輸出之結果也可發現，setosa之正確預測率爲100%，versicolor及 virginica之正確預測率也高達96%及98%。整體而言，正確預測率亦高達98%。因此以線性區別分析做爲預測工具是可行的。

分類結果

		species	預測的各組成員 setosa	預測的各組成員 versicolor	預測的各組成員 virginica	總和
原始的	個數	setosa	50	0	0	50
		versicolor	0	48	2	50
		virginica	0	1	49	50
	%	setosa	100	0	0	100
		versicolor	0	96	4	100
		virginica	0	2	98	100

[a] 98.0% 個原始組別觀察值已正確分類。

由於SPSS中並無二次區別分析之副程式，若在Box's M檢定中棄卻虛無假設時，則可以下列程式進行二次區別分析。

```
SET MXLOOP =99999.
MATRIX.
GET DATA/VARIABLES=X1 to XP/MISSING=OMIT.
GET Group/VARIABLES=G/MISSING=OMIT.
COMPUTE GMAX=CMAX(Group).
COMPUTE GMIN=CMIN(Group).
COMPPUTE NR=NROW(DATA).
COMPPUTE NC=NCOL(DATA).
COMPPUTE RN=MAKE(GMAX-GMIN+1,1,1).
COMPPUTE SSIGMA=MAKE(NC*(GMAX-GMIN+1),NC,0).
COMPUTE PI=4*ARTAN(1).
COMPPUTE QDA=MAKE(NR, (GMAX-GMIN)+1,-999).
COMPPUTE QDAPOST=MAKE(NR, 2*(GMAX-GMIN)+3,-999).
COMPUTE TEMP2=0.
LOOP I=GMIN TO GMAX.
COMPUTE TEMP1=0.
LOOP J=1 TO NR.
DO IF (Group(J)=I).
COMPUTE TEMP1=TEMP1+1.
END IF.
END LOOP.
COMPUTE TEMPDATA=MAKE(TEMP1,NC,999).
LOOP K=1 TO TEMP1.
COMPUTE TEMPDATA(K,:)=DATA(K+TEMP2,:).
END LOOP.
COMPUTE PG=NROW(TEMPDATA)/NR.
COMPUTE TEMP2=TEMP1+TEMP2.
COMPUTE MEAN=T(CSUM(TEMPDATA)/NROW(TEMPDATA)).
COMPUTE SIGMA=(SSCP(TEMPDATA)-NROW(TEMPDATA)*
MEAN*T(MEAN))/(NROW(TEMPDATA)-1).
PRINT I.
PRINT SIGMA.
LOOP V=1 TO NR.
COMPUTE XX=T(DATA(V,:)).
COMPUTE QD=TEMP1/NR*(2*PI)&**(-1*NC/2)*DET(SIGMA)&**(-0.5)*
EXP(-0.5*T(XX-MEAN)*INV(SIGMA)*(XX-MEAN)).
COMPUTE QDA(V,I)=LN(QD)+(NC/2)*LN(2*PI).
```

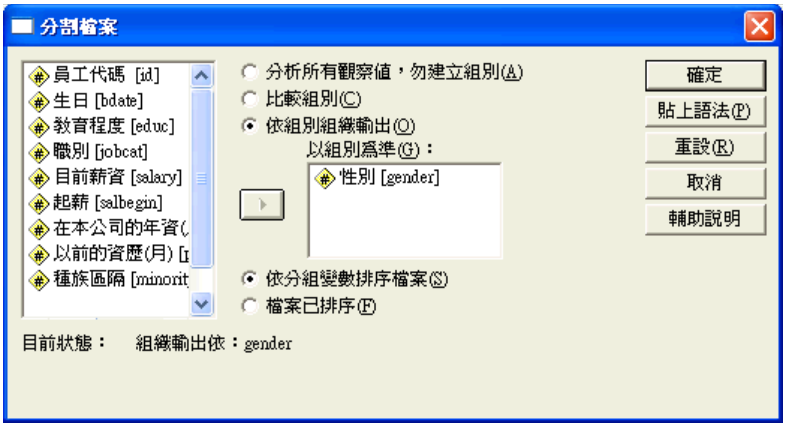
```
COMPUTE QDAPOST(V,I)=LN(QD)+(NC/2)*LN(2*PI).
END LOOP.
END LOOP.
LOOP W= 1 TO GMAX-GMIN+1.
LOOP X=1 TO NR.
COMPUTE QDAPOST(X,GMAX-GMIN+1+W)=QDA(X,W)/RSUM(QDA(X,:)).
END LOOP.
END LOOP.
LOOP X=1 TO NR.
LOOP GPRED= 1 TO GMAX-GMIN+1.
DO IF (QDAPOST(X,GPRED)= RMAX(QDA(X,:))).
COMPUTE QDAPOST(X,2*(GMAX-GMIN)+3)=GPRED.
END IF.
END LOOP.
END LOOP.
PRINT QDAPOST.
END MATRIX.
```

附錄 A

整理資料檔案

A-1 分割檔案

『分割檔案』可以根據某個類別變數，將資料劃分成不同的組別並進行分析。所以，當我們想依據某一類別變數分組作特定一項分析，此時我們就可以利用SPSS裡『分割檔案』的功能。以Employee data.sav爲例，若想針對不同的性別分別利用迴歸分析探討教育程度對起薪的影響，則利用分割檔案功能，依照性別（男：gender=1，女：gender=0）將資料跟割成兩個群。



接下來進行迴歸分析，你會發現SPSS將依照不同的性別分別進行迴歸分析並輸出報表。分割檔案的選項中有『比較組別』及『依組別組織輸出』兩種。

1 比較組別：將分割組別報表排列在一起，方便比較

模式摘要

性別	模式	R	R 平方	調過後的 R 平方	估計的標準誤
女	1	0.474	0.225	0.221	2,590.61
男	1	0.625	0.391	0.388	7,126.95

[a] 預測變數：(常數), 教育程度

2 依組別組織輸出：依照組別將報表分別輸出

模式摘要

模式	R	R平方	調過後的R平方	估計的標準誤
1	0.536	0.287	0.284	6,397.18

[a]預測變數：(常數), 教育程度
[b]性別 = 女

模式摘要

模式	R	R平方	調過後的R平方	估計的標準誤
1	0.646	0.417	0.415	14,918.32

[a] 預測變數：(常數), 教育程度
[b] 性別 = 男

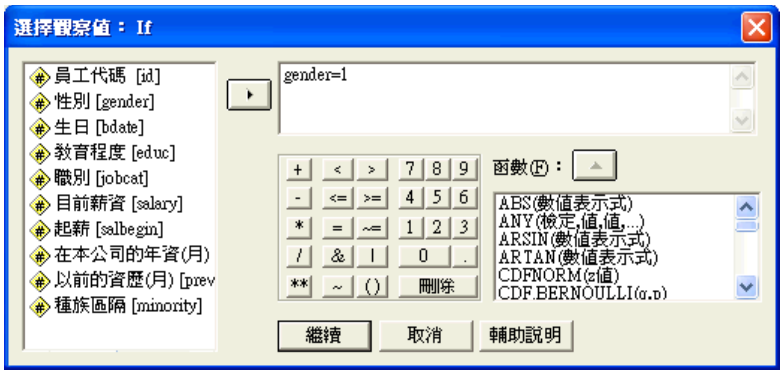
必須注意的是，在SPSS中最多可以指定八個分組變數；如果定義了分割，則它會在以後的所有統計分析中起作用，直到你重新定義分割方式爲止。

A-2 選擇觀察值

當我們只想針對類別變數中的某一類別作特定一項分析，此時我們就可以利用SPSS裡『選擇觀察值』所提供的方法，來選擇觀察值子集。以Employee data.sav爲例，若只想針對男性作迴歸分析探討教育程度對起薪的影響，利用『選擇觀察值』功能來設定資料選取準則。



接下來，在準則輸入方塊中輸入『gender=’m’』。

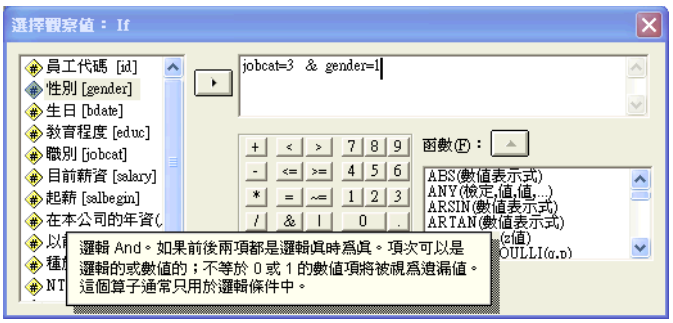


執行之後，你會發現資料視窗的最右邊列號有些有斜線，有些則沒有。有斜線的表示未被選取的資料。

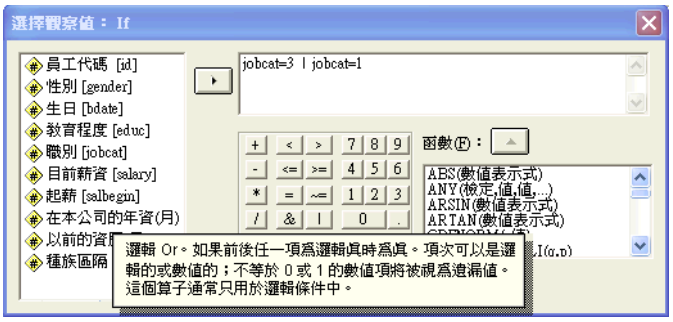
	id	g	bdate	educ	jobcat	salary	salbegin	jobtime	prevexp	minority
209	462	女	10/18/63	16	管理人員	$34,410	$19,500	65	79	否
210	463	女	********	15	職員	$20,700	$14,250	65	241	否
211	466	女	06/15/48	12	職員	$23,400	$13,500	64	198	否
212	467	女	08/18/67	16	職員	$32,850	$19,500	64	20	否
213	468	女	11/28/65	16	管理人員	$55,750	$19,980	64	36	否
214	469	女	06/01/64	15	職員	$25,200	$13,950	64	57	否
215	473	女	********	12	職員	$21,450	$12,750	63	139	否
216	474	女	11/05/68	12	職員	$29,400	$14,250	63	9	否
217	1	男	02/03/52	15	管理人員	$57,000	$27,000	98	144	否
218	2	男	05/23/58	16	職員	$40,200	$18,750	98	36	否
219	5	男	02/09/55	15	職員	$45,000	$21,000	98	138	否
220	6	男	08/22/58	15	職員	$32,100	$13,500	98	67	否
221	7	男	04/26/56	15	職員	$36,000	$18,750	98	114	否
222	12	男	01/11/66	8	職員	$28,350	$12,000	98	26	是
223	13	男	07/17/60	15	職員	$27,750	$14,250	98	34	是
224	15	男	08/29/62	12	職員	$27,300	$13,500	97	66	否
225	16	男	11/17/64	12	職員	$40,800	$15,000	97	24	否
226	17	男	07/18/62	15	職員	$46,000	$14,250	97	48	否
227	18	男	03/20/56	16	管理人員	$103,750	$27,510	97	70	否
228	19	男	08/19/62	12	職員	$42,300	$14,250	97	103	否
229	22	男	09/24/40	12	職員	$21,750	$12,750	97	315	是
230	26	男	11/08/66	15	職員	$31,050	$12,600	96	14	否

當然，有時候會需要更複雜的選取準則，SPSS也提供條件運算式。大部分的條件運算式都可以使用計算機面板上得六個相關算子及邏輯變數，例如：>、<、<=、>=、=、~=、&、|以及~等。茲舉例如下：

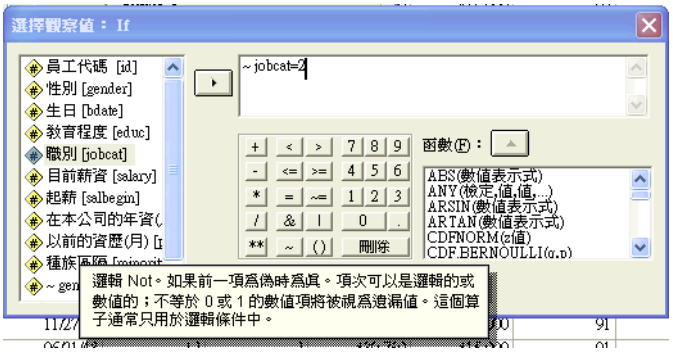
1 And：『&』，例如：輸入『jobcat=3 & gender='m'』表示選取性別爲男性且職稱爲管理主管的資料



2 Or：『|』，例如：輸入『jobcat =1 | jobcat =3』表示選職稱爲職員或管理主管的資料

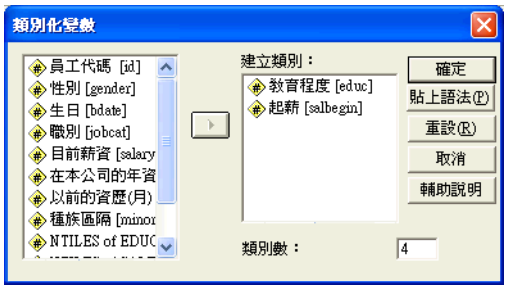


3 Not：『～』，例如：輸入『～ jobcat=2』表示選職稱不是保全人員的資料



A-3 類別化變數

『類別化變數』會將連續數值資料轉換爲非連續類別變數並建立新變數。資料會依據百分位數組別分類，每個組別包含數目近似的觀察值。例如，將資料分爲4的組別會指定1給低於25百分位數、指定2給介於25和50百分位數、指定3給的介於50和75百分位數，並指定4給超過75百分位數的觀察值。以Employee data.sav爲例，若想了解教育程度與起薪是否有關，可將上述兩變數以類別化資料各分爲4組，在以卡分檢定分析。



NTILES of EDUC * NTILES of SALBEGIN 交叉表

		NTILES of SALBEGIN				總和
		1	2	3	4	
NTILES of EDUC	1	19	16	6	4	45
	2	87	66	53	18	224
	3	3	8	16	7	34
	4	7	30	67	67	171
總和		116	120	142	96	474

卡方檢定

	數值	自由度	漸近顯著性 (雙尾)
Pearson卡方	353.027	9	0.000
概似比	348.332	9	0.000
線性對線性的關連	200.416	1	0.000
有效觀察值的個數	474		

A-4 計算

『計算』可以根據其他變數來作數值轉換或計算成某個變數的數值，大部分的運算式都可以使用計算機面板或使用內建函數完成。以Employee data.sav爲例，若想將薪資轉換爲以千元爲單位，首先在目標變數中輸入新變數名稱Nsalary，接下來在數值運算式中輸入salary/1000即可產生nsalary這個新變數。



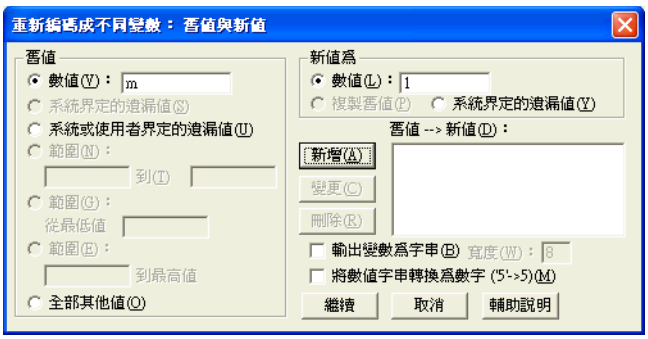
A-5 重新編碼

『重新編碼』的功能在於將變數既有的數值重新設定。當使用者遇到變數的數值必需進行轉換時，可使用此功能。 重新編碼可以將資料輸出爲成不同變數（不會將原來的資料取代置換），也可輸出成相同變數（會將原來的資料置換），是個人需要而定。以Employee data.sav爲例，由於原始資料性別變數屬性爲文字，男：gender=m、女：gender=f。若想將他轉變爲數值資料的新變數，男：sex=1、女：sex=0，則可以下列步驟進行轉換：

Step 1. 將gender輸入數值變數，新變數名稱中鍵入sex，並按變更。



Step 2. 分別輸入舊值=m及新值=1，並按新增；同樣的方法，分別輸入舊值=f及新值=0。



執行之後，你會發現資料視窗會多出一新變數sex。

附錄 B

Iris資料檔

Edgar Anderson's Iris Data: This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.

Id	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa
17	5.4	3.9	1.3	0.4	setosa
18	5.1	3.5	1.4	0.3	setosa
19	5.7	3.8	1.7	0.3	setosa
20	5.1	3.8	1.5	0.3	setosa
21	5.4	3.4	1.7	0.2	setosa
22	5.1	3.7	1.5	0.4	setosa
23	4.6	3.6	1.0	0.2	setosa
24	5.1	3.3	1.7	0.5	setosa
25	4.8	3.4	1.9	0.2	setosa
26	5.0	3.0	1.6	0.2	setosa
27	5.0	3.4	1.6	0.4	setosa
28	5.2	3.5	1.5	0.2	setosa
29	5.2	3.4	1.4	0.2	setosa
30	4.7	3.2	1.6	0.2	setosa
31	4.8	3.1	1.6	0.2	setosa
32	5.4	3.4	1.5	0.4	setosa
33	5.2	4.1	1.5	0.1	setosa
34	5.5	4.2	1.4	0.2	setosa
35	4.9	3.1	1.5	0.2	setosa
36	5.0	3.2	1.2	0.2	setosa
37	5.5	3.5	1.3	0.2	setosa
38	4.9	3.6	1.4	0.1	setosa
39	4.4	3.0	1.3	0.2	setosa
40	5.1	3.4	1.5	0.2	setosa
41	5.0	3.5	1.3	0.3	setosa
42	4.5	2.3	1.3	0.3	setosa

續街下頁···

Id	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
43	4.4	3.2	1.3	0.2	setosa
44	5.0	3.5	1.6	0.6	setosa
45	5.1	3.8	1.9	0.4	setosa
46	4.8	3.0	1.4	0.3	setosa
47	5.1	3.8	1.6	0.2	setosa
48	4.6	3.2	1.4	0.2	setosa
49	5.3	3.7	1.5	0.2	setosa
50	5.0	3.3	1.4	0.2	setosa
51	7.0	3.2	4.7	1.4	versicolor
52	6.4	3.2	4.5	1.5	versicolor
53	6.9	3.1	4.9	1.5	versicolor
54	5.5	2.3	4.0	1.3	versicolor
55	6.5	2.8	4.6	1.5	versicolor
56	5.7	2.8	4.5	1.3	versicolor
57	6.3	3.3	4.7	1.6	versicolor
58	4.9	2.4	3.3	1.0	versicolor
59	6.6	2.9	4.6	1.3	versicolor
60	5.2	2.7	3.9	1.4	versicolor
61	5.0	2.0	3.5	1.0	versicolor
62	5.9	3.0	4.2	1.5	versicolor
63	6.0	2.2	4.0	1.0	versicolor
64	6.1	2.9	4.7	1.4	versicolor
65	5.6	2.9	3.6	1.3	versicolor
66	6.7	3.1	4.4	1.4	versicolor
67	5.6	3.0	4.5	1.5	versicolor
68	5.8	2.7	4.1	1.0	versicolor
69	6.2	2.2	4.5	1.5	versicolor
70	5.6	2.5	3.9	1.1	versicolor
71	5.9	3.2	4.8	1.8	versicolor
72	6.1	2.8	4.0	1.3	versicolor
73	6.3	2.5	4.9	1.5	versicolor
74	6.1	2.8	4.7	1.2	versicolor
75	6.4	2.9	4.3	1.3	versicolor
76	6.6	3.0	4.4	1.4	versicolor
77	6.8	2.8	4.8	1.4	versicolor
78	6.7	3.0	5.0	1.7	versicolor
79	6.0	2.9	4.5	1.5	versicolor
80	5.7	2.6	3.5	1.0	versicolor
81	5.5	2.4	3.8	1.1	versicolor
82	5.5	2.4	3.7	1.0	versicolor
83	5.8	2.7	3.9	1.2	versicolor
84	6.0	2.7	5.1	1.6	versicolor
85	5.4	3.0	4.5	1.5	versicolor
86	6.0	3.4	4.5	1.6	versicolor
87	6.7	3.1	4.7	1.5	versicolor
88	6.3	2.3	4.4	1.3	versicolor
89	5.6	3.0	4.1	1.3	versicolor
90	5.5	2.5	4.0	1.3	versicolor
91	5.5	2.6	4.4	1.2	versicolor
92	6.1	3.0	4.6	1.4	versicolor
93	5.8	2.6	4.0	1.2	versicolor
94	5.0	2.3	3.3	1.0	versicolor
95	5.6	2.7	4.2	1.3	versicolor
96	5.7	3.0	4.2	1.2	versicolor
97	5.7	2.9	4.2	1.3	versicolor
98	6.2	2.9	4.3	1.3	versicolor
99	5.1	2.5	3.0	1.1	versicolor
100	5.7	2.8	4.1	1.3	versicolor
101	6.3	3.3	6.0	2.5	virginica
102	5.8	2.7	5.1	1.9	virginica
103	7.1	3.0	5.9	2.1	virginica
104	6.3	2.9	5.6	1.8	virginica
105	6.5	3.0	5.8	2.2	virginica

續街下頁···

Id	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
106	7.6	3.0	6.6	2.1	virginica
107	4.9	2.5	4.5	1.7	virginica
108	7.3	2.9	6.3	1.8	virginica
109	6.7	2.5	5.8	1.8	virginica
110	7.2	3.6	6.1	2.5	virginica
111	6.5	3.2	5.1	2.0	virginica
112	6.4	2.7	5.3	1.9	virginica
113	6.8	3.0	5.5	2.1	virginica
114	5.7	2.5	5.0	2.0	virginica
115	5.8	2.8	5.1	2.4	virginica
116	6.4	3.2	5.3	2.3	virginica
117	6.5	3.0	5.5	1.8	virginica
118	7.7	3.8	6.7	2.2	virginica
119	7.7	2.6	6.9	2.3	virginica
120	6.0	2.2	5.0	1.5	virginica
121	6.9	3.2	5.7	2.3	virginica
122	5.6	2.8	4.9	2.0	virginica
123	7.7	2.8	6.7	2.0	virginica
124	6.3	2.7	4.9	1.8	virginica
125	6.7	3.3	5.7	2.1	virginica
126	7.2	3.2	6.0	1.8	virginica
127	6.2	2.8	4.8	1.8	virginica
128	6.1	3.0	4.9	1.8	virginica
129	6.4	2.8	5.6	2.1	virginica
130	7.2	3.0	5.8	1.6	virginica
131	7.4	2.8	6.1	1.9	virginica
132	7.9	3.8	6.4	2.0	virginica
133	6.4	2.8	5.6	2.2	virginica
134	6.3	2.8	5.1	1.5	virginica
135	6.1	2.6	5.6	1.4	virginica
136	7.7	3.0	6.1	2.3	virginica
137	6.3	3.4	5.6	2.4	virginica
138	6.4	3.1	5.5	1.8	virginica
139	6.0	3.0	4.8	1.8	virginica
140	6.9	3.1	5.4	2.1	virginica
141	6.7	3.1	5.6	2.4	virginica
142	6.9	3.1	5.1	2.3	virginica
143	5.8	2.7	5.1	1.9	virginica
144	6.8	3.2	5.9	2.3	virginica
145	6.7	3.3	5.7	2.5	virginica
146	6.7	3.0	5.2	2.3	virginica
147	6.3	2.5	5.0	1.9	virginica
148	6.5	3.0	5.2	2.0	virginica
149	6.2	3.4	5.4	2.3	virginica
150	5.9	3.0	5.1	1.8	virginica

附録 C

機率分配表

標準常態分配表; $P(Z < z)$

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.10	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.20	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.30	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.40	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.50	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.60	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.70	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.80	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.90	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.00	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.10	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.20	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.30	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.40	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.50	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.60	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.70	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.80	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.90	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.00	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.10	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.20	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.30	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.40	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.50	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.60	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.70	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.80	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.90	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.00	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.10	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.20	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.30	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.40	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.50	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998

Student-t 分配表; $P(T > t)$

	α					
d.f	0.25	0.1	0.05	0.025	0.01	0.005
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707
7	0.711	1.415	1.895	2.365	2.998	3.499
8	0.706	1.397	1.860	2.306	2.896	3.355
9	0.703	1.383	1.833	2.262	2.821	3.250
10	0.700	1.372	1.812	2.228	2.764	3.169
11	0.697	1.363	1.796	2.201	2.718	3.106
12	0.695	1.356	1.782	2.179	2.681	3.055
13	0.694	1.350	1.771	2.160	2.650	3.012
14	0.692	1.345	1.761	2.145	2.624	2.977
15	0.691	1.341	1.753	2.131	2.602	2.947
16	0.690	1.337	1.746	2.120	2.583	2.921
17	0.689	1.333	1.740	2.110	2.567	2.898
18	0.688	1.330	1.734	2.101	2.552	2.878
19	0.688	1.328	1.729	2.093	2.539	2.861
20	0.687	1.325	1.725	2.086	2.528	2.845
21	0.686	1.323	1.721	2.080	2.518	2.831
22	0.686	1.321	1.717	2.074	2.508	2.819
23	0.685	1.319	1.714	2.069	2.500	2.807
24	0.685	1.318	1.711	2.064	2.492	2.797
25	0.684	1.316	1.708	2.060	2.485	2.787
26	0.684	1.315	1.706	2.056	2.479	2.779
27	0.684	1.314	1.703	2.052	2.473	2.771
28	0.683	1.313	1.701	2.048	2.467	2.763
29	0.683	1.311	1.699	2.045	2.462	2.756
30	0.683	1.310	1.697	2.042	2.457	2.750
40	0.681	1.303	1.684	2.021	2.423	2.704
60	0.679	1.296	1.671	2.000	2.390	2.660
120	0.677	1.289	1.658	1.980	2.358	2.617
∞	0.674	1.282	1.645	1.960	2.326	2.576

χ^2分配表; $P(\chi^2 > x)$

	α							
d.f	0.99	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	0.00	0.00	0.02	2.71	3.84	5.02	6.63	7.88
2	0.02	0.10	0.21	4.61	5.99	7.38	9.21	10.60
3	0.11	0.35	0.58	6.25	7.81	9.35	11.34	12.84
4	0.30	0.71	1.06	7.78	9.49	11.14	13.28	14.86
5	0.55	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	0.87	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	1.24	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.65	2.73	3.49	13.36	15.51	17.53	20.09	21.95
9	2.09	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.56	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	3.05	4.57	5.58	17.28	19.68	21.92	24.72	26.76
12	3.57	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	4.11	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.66	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	5.23	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.81	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	6.41	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	7.01	9.39	10.86	25.99	28.87	31.53	34.81	37.16
19	7.63	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	8.26	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.90	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	9.54	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	10.20	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	10.86	13.85	15.66	33.20	36.42	39.36	42.98	45.56
25	11.52	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	12.20	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	12.88	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	13.56	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	14.26	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	14.95	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	22.16	26.51	29.05	51.81	55.76	59.34	63.69	66.77
50	29.71	34.76	37.69	63.17	67.50	71.42	76.15	79.49
60	37.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
70	45.44	51.74	55.33	85.53	90.53	95.02	100.43	104.21
80	53.54	60.39	64.28	96.58	101.88	106.63	112.33	116.32
90	61.75	69.13	73.29	107.57	113.15	118.14	124.12	128.30
100	70.06	77.93	82.36	118.50	124.34	129.56	135.81	140.17

索引

α風險, 22
β風險, 22
$\mathcal{F}$分配, 6
t分配, 6

Bartlett's test of sphericity, 102
Box'M test, 145
Box-Cox transformation, 74

Cook's Distance, 79

Fisher's exact test, 42

KMO量測值, 115
Kruskal-Wallis test, 60

Lagrange乘數法, 100
Levene's test, 30

P-value, 23

Q-Q plot, 9

SMC, 114

Yates' correction for continuity, 42

中位數, 2
中央極限定理, 7
主成份分分數, 100
主成份分析, 99
主成份負荷, 101
二次區別函數, 141
二次區別分數, 141
信賴上限, 21
信賴下限, 21
信賴區間, 21
信賴水準, 21

假設檢定, 22
共同性, 101
共線性, 89
典型區別分析, 142
典型權重, 125
典型相關係數, 125
典型負荷量, 126
判定係數, 68
區間估計, 21
區集設計, 48
十分位數, 2
卡方分配, 5
危險域，棄卻域, 22
向前選取法, 91
向後選取法, 91
四分位差, 3
四分位數, 2
四次方最大值法, 115
因子, 47
因子水準, 47
因素分析，主因子分析法 , 113

因素分析，主成分分析法, 113
因素分析，最大概似法, 113
因素負荷量, 111
型I誤差, 22
型II誤差, 22
多元迴歸分析, 83
多重比較, 59
大數法則, 7
學生化殘差, 80
完全隨機化設計, 48
容忍度, 91
對立假設, 22
常態分配, 5
弱大數法則, 7
強大數法則, 7

抽樣分配, 2
散佈圖, 9
斜交轉軸, 115
普通最小平方法, 64
最大變異法, 115
期望誤歸成本, 136
柴比雪夫定理, 4
條件指標, 91
概似比檢定統計量, 41
槓桿值, 80
標準分數, 4
樣本變異數, 3
母體參數, 1
測量尺度, 4
獨立性檢定, 40
百分位數, 3
盒形圖, 8
直交轉軸, 115
直方圖, 8
相等最大值法, 115
衆數, 3
簡單假設, 22
簡單線性迴歸分析, 63
統計假設, 22
統計量, 1
經驗法則, 4
線性區別函數, 142
線性區別分數, 142
總誤歸率, 136

脊迴歸, 90
處理, 47
虛無假設, 22
複合假設, 22
調整後判定係數, 88
變異係數, 4
變異數分析, 47
變異數膨脹因素, 91

逐步迴歸法, 91
適合度檢定, 37
重疊指標, 127
陡坡圖, 101
隨機樣本, 1
離中平均差, 3

離均平均差, 3
離群値, 79
顯著性檢定, 24
顯著水準, 23
齊一性檢定, 45

國家圖書館出版品預行編目資料

量化研究與統計分析／戴忠淵、孔方正編撰 – 初版

高雄市：2013（民102）

面：19.7 × 27.2公分

ISBN 978 – 986 – 80629 – 6 – 2　　（平裝）

量化研究與統計分析

發 行 人： 樹德科技大學

編 撰 者： 戴忠淵、孔方正

出 版 者： 宏冠出版社

印刷裝訂： 宏冠數位印刷 07-5315351
http://www.pod.com.tw

出版日期： 中華民國一〇二年九月出版

※本文章非經書面同意，不得轉載※