CABLE 2006 教育訓練
SAS 程式使用

講義編寫：吳文琪

2006/8/9

目　　錄

壹、基本概念

一、 資料分析步驟

（一）瞭解資料的結構

一個資料庫通常會有三個部分，一是問卷，二是資料檔案，三是譯碼簿(Coding Book)。在資料分析時，雖進入統計軟體分析的僅有資料檔案，但為要瞭解每個變項的意義，仍需配合譯碼簿和問卷。

為了使分析有依據，通常會有分析架構，配合不同變項類型，採取不同的分析方法。

（二）將資料輸入及輸出

一般資料可能會有幾種格式：

1. 直接是該統計軟體的格式
2. 文字檔案
3. excel 等資料庫的格式

通常針對不同的格式，匯入資料的方法也不同。

而針對不同的資料分析方法，也有不同的統計軟體可以使用。目前常聽到的統計軟體包括 SAS、SPSS、SPLUS、MPLUS、STATA、LISREL 等，此處以 SAS 做為主要統計分析軟體。

（三）資料清理

資料清理亦可分為幾個小步驟：

1. 釐清研究目的
2. 選取需要的變項
3. 瞭解這些變項的分佈
4. 觀察遺漏值和不合理值
5. 配合架構，進行變項的運算，例如加總、反向計分或跳答重計分等。
6. 整理出配合研究目的及架構的資料檔案。

（四）進入分析

根據變項的型態及不同的分析目的，必須採取不同的資料分析方法，舉例如下。

目的 型態	單變項描述	雙變項關係		多變項關係
		連續	類別	
連續型	平均值、標準差、最大值、最小值	相關係數		線性複回歸
類別型	百分率	T 檢定、ANOVA	卡方檢定	邏輯斯複回歸

另外還有：

1. 無母數的資料分析：樣本數很小的時候，必須採用的分析方法。
2. 重複測量的資料分析：相依樣本（非獨立的樣本）或長期追蹤時會用到的分析方法。
3. 多變量的資料分析：因素分析、集群分析、結構方程模式、階層迴歸等。

統計分析方法琳瑯滿目、包羅萬象，在這裡，僅以非常常用的方法，配合 SAS 統計軟體的使用，進行介紹。

二、 名詞定義

（一）變項

1. 或稱變數，指會隨觀察對象（受訪者）之不同，而有不同觀察值。
2. 若所收集到的資料，發現所有受訪者的某一變項結果全部相同。此變項或許仍可呈現其單變項描述，但應用性不高。
3. 變項若細分可分為四種類型，nominal、ordinal、interval、ratio，但在實際分析時，可粗分為類別變項和連續變項兩類。在行為分析時，主要是以 interval 的變項作為連續變項。

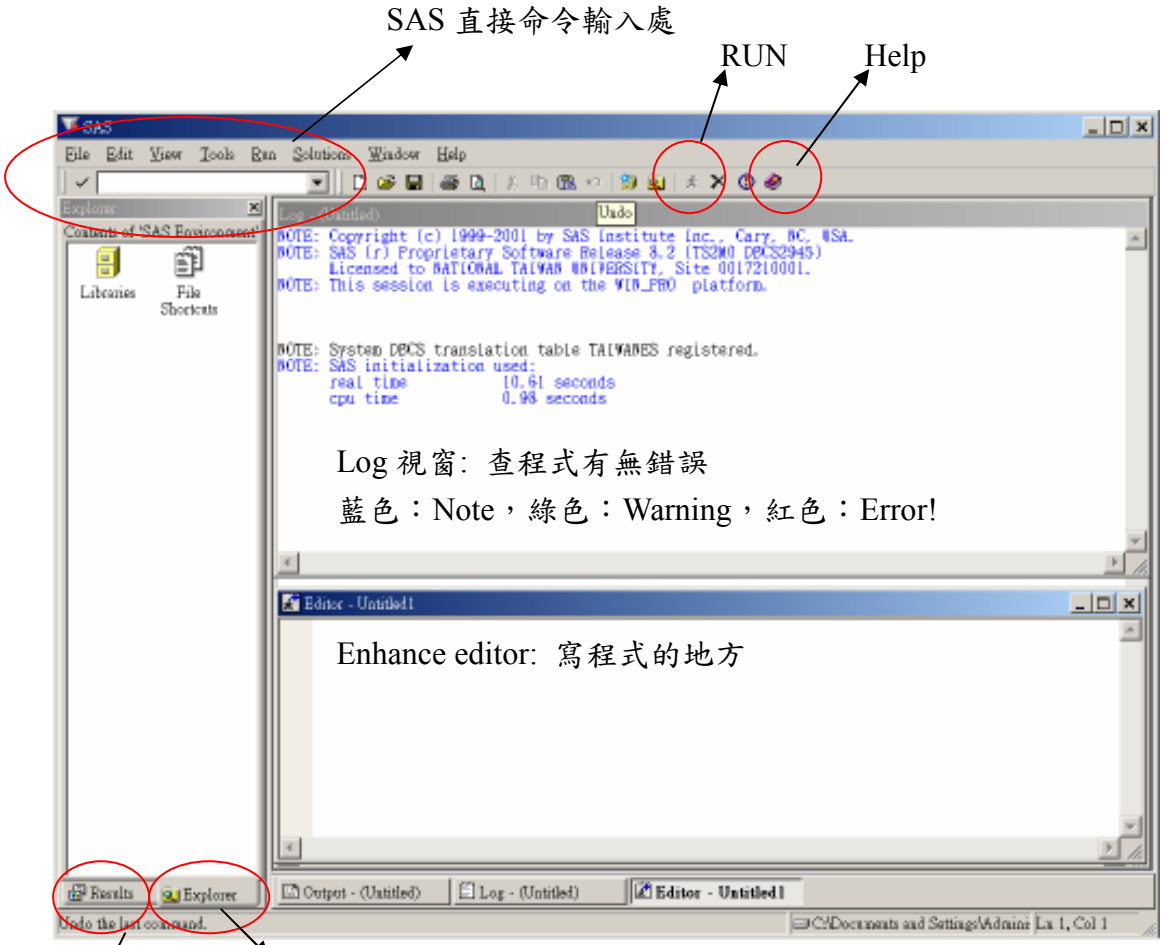
（二）概念

1. 指一組類似的變項，用來表達一個意涵，通常有某些理論依據。
2. 概念(concept)是理論的主要成份，通常指一抽象意念，很難直接觀察的，例如自我效能、對吸菸的態度等，亦可以說概念是為表達出一某種觀察現象歸納的抽象意思。
3. 概念下可以包含很多變項，變項下又可以包含很多題目。

（三）關係

1. 通常是指兩個變項或概念是否具有某種互動，例如 A 升 B 升，A 升 B 降等。
2. 目前常用的是正相關（互動方向相同）、負相關（互動方向不同），及有關（無法判斷方向）三種。
3. 統計上沒有顯著關係，不表示實際沒有顯著關係。

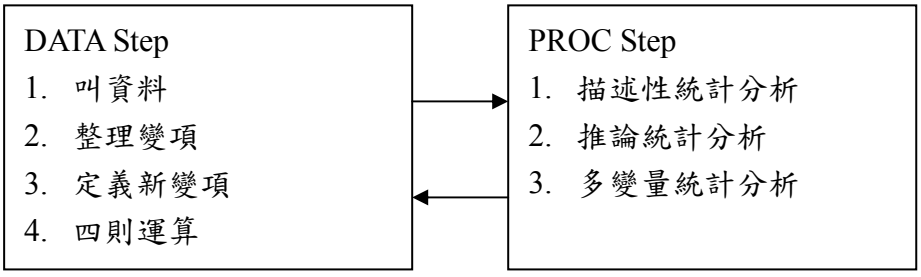
三、 SAS 視窗環境



（一）SAS 程式兩大部分

1. DATA step：資料處理的部分。
2. PROC step：統計分析的部分。

*資料處理一定要在統計分析之前。



（二）執行 SAS 的步驟

1. 進入 SAS 系統
2. 確定分析資料儲存路徑
3. 在 Program Editor 視窗下建立和編輯 SAS 程式
4. 執行程式
5. 在 log 視窗檢視錯誤訊息
6. 在 output 視窗看輸出結果

（三）SAS 的 library 功能

暫存檔和永久檔的差別

1. 暫存檔：SAS 開著時，存在 WORK 資料夾內，SAS 關掉後就消失了。
2. 永久檔：SAS 關掉後仍能繼續存在，需存到研究者自行定義的目錄裡。
3. 故若有「目錄.資料名稱」（如 DATA.ab、XX.yy），則為永久檔，若直接是資料名稱（如 ab、yy），則是存在 WORK 內的暫存檔。

（四）功能鍵及簡易直接命令

1. [F5]：Program editor 視窗
2. [F6]：Log 視窗
3. [F7]：Output 視窗
4. [F3]或[F8]：執行程式
5. [F4]：Recall 最後一次執行的程式，但關 SAS 後再開，不會記憶最後一次程式。
6. clear all：清除所有程式
7. help：呼叫 SAS help
8. insight：呼叫 SAS insight

（五）各種副檔名

1. 「*.sas」：程式的副檔名
2. 「*.log」：log 的副檔名
3. 「*.list」：output 的副檔名
4. 「*.ssd」、「*.sd2」、「sas7bdat」：dataset 的副檔名
5. 以上 1~3 均可用文字處理軟體閱讀，如 WORD 或 NOTEPAD 等。
6. 資料檔命名原則：

(1)名稱中間不可有空白、點、dash 等字樣

(2)名稱中可以有底線

(3)必須以英文字母或底線開頭，不可用數字開頭。

(4)變項名稱不可超過 32 個位元。

（六）小常識

1. 所有程式均需以「;」結尾。

2. 「/*」和「*/」內的整段文字是註解，SAS 不會去執行。「*」表示此行是註解，SAS 也不會去執行。
3. Enhance editor 的程式內，紅色表示錯誤（error），綠色表示註解，藍色表示程式語言，紫紅色表示文字。這些設定均可在 Tools/ options/ enhance editor/ Appearance 中修改。
4. Log 的部分，藍色是註解（Note），綠色是警告（Warning），紅色是錯誤（Error）
5. 產生 MSOffice 可用的 Output 格式：Tools/ options/ preference/ Results/ Create html/ 按自己的喜好勾選。存放在 Work 目錄下，關程式後，結果就會消失。

貳、SAS 的 DATA step 常用程式

一、 系統環境設定（option）

（一）在 DATA step 之前使用的選項

用來設定 SAS output 的畫面，及初步資料擷取。

（二）選項

1. CENTER/NOCENTER：結果是否置中
2. DATE/NODATE：是否要呈現日期
3. PAGENO=n：每次頁碼都從第 n 頁開始
4. NONUMBER：不顯示頁碼
5. LINESIZE/LS=n：一行的字元數
6. PAGESIZE/PS=n：設定每頁的行數
7. FIRSTOBS=n：從第幾筆資料開始抓
8. OBS=n 或 Max：最多抓到第幾筆

（三）程式範例[prog01.sas]

OPTIONS CENTER LS=80 PAGENO=1 NODATE;

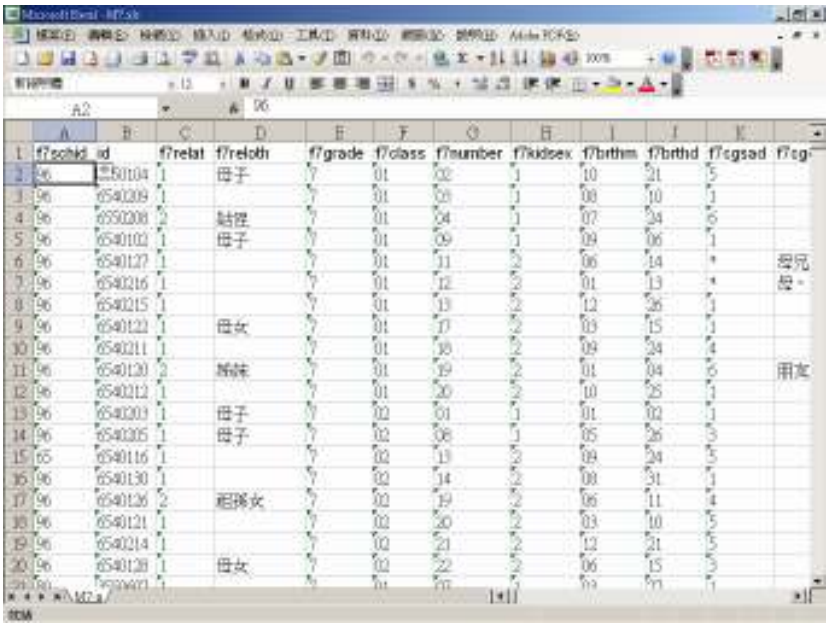
二、 資料輸入及輸出

（一）資料型態

一般資料來源，可分為內部資料和外部資料。內部資料是指經由 SAS 建檔的資料，可由 SAS 直接讀取，分成讀取 SAS 檔案和直接在程式輸入兩種。外部資料是指需要用不同形式叫進來看的資料，目前常用的有 Excel 和文字檔。

（二）Excel 檔

1. 資料樣貌（*.xls）



2. 程式[prog02.sas]

```
/*Import exel data*/
PROC IMPORT OUT= DATA.m7
            DATAFILE=
            "E:\Wen_chi\CABLE\Cable2006EducationTrainning\Data\M7.xls"
            DBMS=EXCEL2000 REPLACE;
            GETNAMES=YES;
RUN;

/*Import dBase data*/
PROC IMPORT OUT= DATA.M7DBF
            DATAFILE=
            "E:\Wen_chi\CABLE\Cable2006EducationTrainning\Data\M7.dbf"
            DBMS=DBF REPLACE;
            GETDELETED=NO;
RUN;
```

3. 使用須知

(1) 點選路徑：File/ import data/ standard data source/ where is the file located?/browse (or type)/ choose SAS destination/ browse (to save program)

(2) 可用上述路徑將資料匯入的程式存起來，以利記錄和方便重複使用。

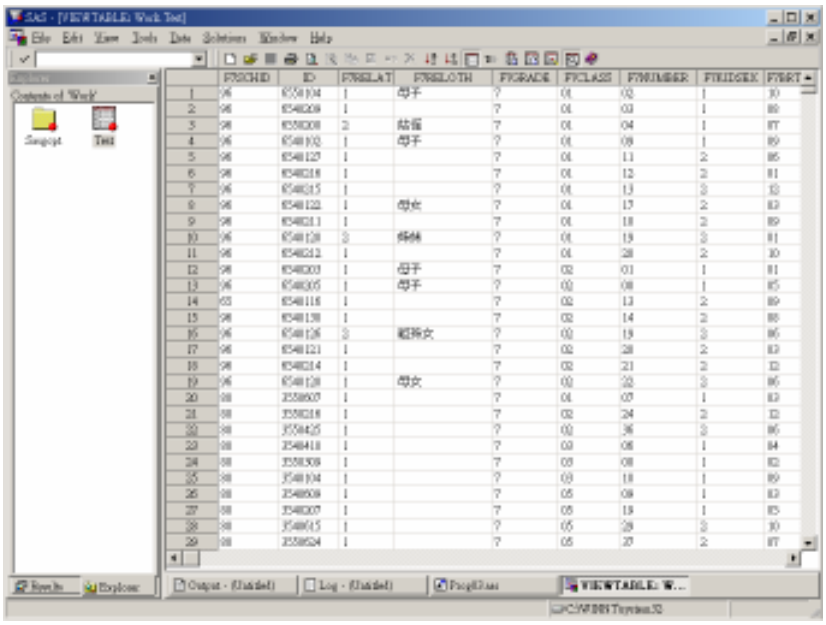
(3) Excel 有時會有版本不相容的問題，可將其轉換為 Dbase4 的檔案（*.dbf），轉換時需將 excel 的資料調成最適欄位。

(4) 建檔時，資料名稱建議用英文，變項名稱中不可有「.」或「-」。

(5) Excel 最長只能放 256 個欄位。

（三）SAS 檔

1. 資料樣貌（*.sas7bdat）（*.ssd）（*.sd2）



2. 程式[prog03.sas]

```
/*Import sas data and typing data*/
LIBNAME train 'E:\Wen_chi\CABLE\Cable2006EducationTrainning\Data';
DATA chrt2k;
SET train.chrt2k04c;
RUN;
```

3. 程式註解

(1) **LIBNAME** 定義檔案存放位置的路徑。

(2) **SET** 叫資料進來。

(3) **DATA** 存資料。

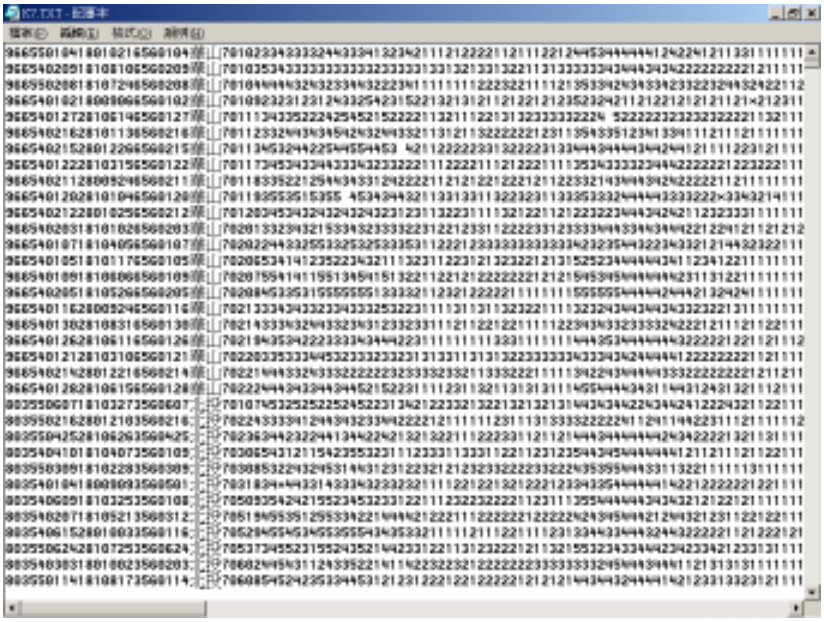
4. 使用須知

(1) 除用上述程式外，可直接用 SAS insight 點選來看，另外，若有灌 SAS viewer，則可直接點選打開 SAS 資料檔。

(2) 只要有 **LIBNAME**，就可以用 SAS Library 點選看資料。

（四）文字檔

1. 資料樣貌（*.txt）



2. 程式[prog04.sas]

```
LIBNAME data 'E:\Wen_chi\CABLE\Cable2006EducationTrainning\Data';
DATA data.k7;
INFILE 'E:\Wen_chi\CABLE\Cable2006EducationTrainning\Data\K7.txt' LRECL=411;
INPUT
      y7schid $1-2 id $3-9 y7kidsex $10-10 y7brthy $11-12 y7brthm $13-14
      y7brthd $15-16 y7lstsch $17-18 y7lstgrd $19-19 y7lstcls $20-21 y7lstnmb $22-23
      y7schnm $24-27 (其他省略);
RUN;
```

3. 程式註解

(1) **INFLE** 表明外部資料之文字檔存放地點。**LRECL** 是資料的欄位數。(通常需要看 codebook 才會知道)

(2) **INPUT** 給定資料格式：「變項名稱＋格式＋欄位座標」是一組程式，格式的部分，數字不用特別表示，文字需用「$」表示。

(3) **INFILE＋INPUT** 是一組特定程式，需一起使用。

4. 使用須知

(1) 文字檔鍵入時，可用「.」表示 missing，若是有固定格式的輸入，也可用「 」空白當作 missing。

（五）直接鍵入

1. 資料樣貌

直接在程式中鍵入資料。

2. 程式[prog03.sas]

(1) 每4個觀測值構成一筆記錄

```
DATA double;
```

```
  INPUT group $ score @@;
DATALINES;
1 67 1 87 1 98 1 50
2 44 2 67 2 49 2 83
3 99 3 98 3 88 3 69
;
RUN;
```

(2) 社區和居民的資訊混在一起

```
DATA single;
  INPUT type $ 1. @;
  IF type='c' THEN INPUT name $ pop;
  ELSE IF type='r' THEN INPUT income tax;
CARDS;
c monroe 8000
c green 15340
r 19000 3520
r 65000 20000
;
RUN;
```

3. 程式註解
   (1) 用「@@」表示連續讀取，用「@」表示一列一列讀。
   (2) 用 **DATALINES** 或 **CARDS** 叫資料。
   (3) 用用 **INPUT** 設定變項名稱和格式。
   (4) 配合 **IF….THEN; ELSE IF…..THEN….**;抽取不同列的資料。
4. 使用須知
   (1) 資料較少時，會使用直接鍵入的方式。
   (2) 做簡單運算時，可以使用直接鍵入的方式。

（六）三種讀取資料的方法：[prog03.sas]

	資料樣式	使用須知
簡列讀入法 List input	資料型態很整齊	1. 變項和變項之間需要有空格相隔。 2. 文字變項內不可有空格。
欄位讀入法 Column input	資料連接在一起時	1. 變項間不需要有空格。 2. 文字變項內可以有空格
格式讀入法 Formatted input	資料很亂的時候	粉複雜的格式也可以用。

1. 簡列讀入法 List input

```
DATA test01;
      INPUT name $ sex $ age height weight;
DATALINES;
Alice f 14 168 50
May f 16 172 60
Paul m . 180 70
;
RUN;
```

2. 欄位讀入法 Column input

```
DATA test01;
      INPUT name $ 1-6 sex $ 7-8 age 9-11
            height 12-15 weight 16-17;
DATALINES;
Alice  f 14 16850
May wu f 16 17260
Paul   m .  18070
;
RUN;
```

3. 格式讀入法 Formatted input

```
DATA test01;
      INPUT #1 name $ 6.0 / @ 1 sex $ /@ 1 age /@ 1 height weight;
DATALINES;
Alice
  f
   14
     168 50
May wu
   f
  16
    172 60
Paul
  m
 .
 180 70
;
RUN;
```

指令	意義
@n	跳到第 n 欄或第 n 行
#n	跳到第 n 筆記錄或第 n 列
+n	略過 n 欄或 n 行
/	略過至下一筆記錄、換列

（七）資料輸出：

1. 用 Libname 的方式，將 SAS 資料存成永久檔。

2. 用點選及配合將程式存起來的方式，將 SAS 資料輸出成各種不同的檔案。File/ Export/ Choose the source of SAS data set/ 選好後按照接下來的選單選擇要輸出的格式，最後選擇處存位置及輸入檔名即可。

[prog02.sas]

```
PROC EXPORT DATA= WORK.SINGLE
            OUTFILE= "E:\Wen_chi\temp\single.txt"
            DBMS=TAB REPLACE;
RUN;
```

三、 產生分析需用的資料檔

（一）選擇或刪除樣本 [prog05.sas]

1. 用途

在很多樣本中，選取或刪除某些特定條件的樣本。

2. 程式

(1)選擇樣本：用**IF**留下樣本

```
DATA boy; SET chrt2k; IF y7kidsex=1; RUN;
```

(2)刪除樣本：用**DELETE**刪除樣本

```
DATA girl; SET chrt2k; IF y7kidsex=1 THEN DELETE; RUN;
```

(3)直接切成兩個資料檔：符合某條件進一個資料檔，符合另一個條件進入另一個資料檔。用**IF**配**OUPUT**存檔。

```
DATA boy girl;
SET chrt2k;
IF y7kidsex=1 THEN OUTPUT boy;
IF y7kidsex=2 THEN OUTPUT girl;
run;
```

（二）選擇或刪除變項

1. 用途

在很多變項中，選擇或刪除特定的變項。

2. 程式

(1) 選擇變項：用 **KEEP** 選擇變項，單獨一行或在 SET 後面均可。

```
DATA test1; SET chrt2k; KEEP id y7lstsch y7kidsex; RUN;
DATA test2; SET chrt2k (KEEP=id y7class y7number); RUN;
```

(2) 刪除變項：用 **DROP** 選擇變項，單獨一行或在 SET 後面均可。

```
DATA test3; SET chrt2k; DROP y7number; RUN;
DATA test4; SET chrt2k (DROP=y7number y7grade); RUN;
```

3. 使用須知

(1) **KEEP** 和 **DROP** 均可放在 **DATA** 或 **SET** 的後面，效果相同，也可以

搭配 **PROC** 的 **PROCEDURE** 使用，甚至可以用 **KEEP** 或 **DROP** 將一個大檔切成幾個不同的小檔。

(2) 例 1：把大檔 cc 變成小檔 aa（去除變項 z）和另一個小檔 bb（僅留變項 x 和 y）：

```
DATA aa (DROP=z)
        bb (KEEP=x y);
 SET cc;
 RUN;
```

(3) 例 2：列印所有變項，但不列印 id。

```
PROC PRINT (DROP=id); RUN;
```

（三）合併資料檔

1. 用途：把兩個以上的小資料，根據某項原則，合併成大資料。
2. 程式

(1) 合併欄位：合併後變項增加，用 **MERGE**。

```
DATA test1; SET test1; PROC SORT; BY id; run;
DATA test2; SET test2; PROC SORT; BY id; run;
DATA Merge1; MERGE test1 test2; BY id; run;
```

(2) 合併樣本：合併後樣本數增加，用 **SET**。

```
DATA sex; SET boy girl; RUN;
```

3. 程式註解：用 **MERGE** 時，小資料需要先經過 **PROC SORT** 排序，且需根據可以某個可以辨認出不同樣本的變項（如 ID）來排序。

四、 邏輯用語

（一）IF..THEN 條件句原件

名稱	符號	名稱	符號
大於	>	和	AND, &
小於	<	或	OR, |
等於	=	非、不是	NOT, ^
不等於	^=, ~=, NE		
大於等於	>=		
小於等於	<=		

（二）根據某條件做一件事

IF….THEN; ex:如果學校代碼(y7lstsch)介於某範圍，則地區屬於台北或新竹。

```
DATA area;
SET test1;
```

```
IF 11<=y7lstsch<=36 THEN area='1=Taipei';
IF 41<=y7lstsch<=66 THEN area='2=Hsinju';
run;
```

（三）根據某條件做很多事

IF….THEN DO; END;

Ex1:如果學校代碼等於某範圍或數字，則地區和學校大小各有其不同的屬性。

```
DATA area2;
SET test1;
IF y7lstsch=11 THEN DO; area='Taipei'; schsize='big'; END;
IF y7lstsch=21 or y7lstsch=22 THEN DO; area='Taipei'; schsize='med'; END;
IF 31<=y7lstsch<=36  THEN DO; area='Taipei'; schsize='sml'; END;
IF y7lstsch=41 THEN DO; area='Hsin'; schsize='big'; END;
IF y7lstsch=51 or y7lstsch=52 THEN DO; area='Hsin'; schsize='med'; END;
IF 61<=y7lstsch<=66 THEN DO; area='Hsin'; schsize='sml'; END;
run;
```

Ex2:需要設定 dummy variable 時，可使用此程式。例如 VAR1 有四類或四個選項，需要做成三個 dummy variable。（將一個變項，變成三個變項）

Dummy Variable / VAR1	D1	D2	D3
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1

```
IF VAR1=1 THEN DO; D1=0; D2=0; D3=0; END;
IF VAR1=2 THEN DO; D1=1; D2=0; D3=0; END;
IF VAR1=3 THEN DO; D1=0; D2=1; D3=0; END;
IF VAR1=4 THEN DO; D1=0; D2=0; D3=1; END;
```

（四）排除上一個條件後，做下一件事

IF….THEN; ELSE IF …..THEN; ELSE….; ex:如果學校代碼是 xx，則學校大小就是 aa；除了 xx 的學校以外，若學校代碼為 yy，則學校大小就是 bb；除了學校代碼是 xx 和 yy 以外的學校，剩下的學校其學校大小均為 cc。

```
DATA area3;
SET test1;
IF y7lstsch=11 or y7lstsch=41 THEN schsize='big';
```

```
ELSE IF 21<=y7lstsch<=22 or 51<=y7lstsch<=52 THEN schsize='med';
ELSE IF 31<=y7lstsch<=36 or 61<=y7lstsch<=66 THEN schsize='sml';
ELSE schsize=.;
RUN;
```

五、 基本運算

（一）四則運算

運算名稱	代號	舉例	說明
加	+	W = x + z ;	
減	-	W = x – z ;	
乘	*	W = x * z ;	
除	/	W = x / z ;	
次方	**	W = x**2 ;或 W = x**3 ;	x 的平方或三次方
反向	-	W= - x ;	

（二）基本函數

1. 數學函數

函數名稱	舉例	說明
絕對值	Y=ABS(x)	
餘數	Y=MOD(10,3)	10 除 3 的餘數
平方根	Y=SQRT(4)	Y=2
e 的 n 次方	Y=EXP(2)	$Y=e^2$
自然對數	Y=LOG(1)	1 取自然對數，所以 Y=0
以 n 為底的對數	Y=LOG10(100)	100 取以 10 為底的對數
四捨五入	Y=ROUND(x, 1) Y=ROUND(x, 0.1) Y=ROUND(x, 0.01)	四捨五入至整數 四捨五入至小數第一位 四捨五入至小數第二位

2. 統計值函數

統計值的函數會跳過 missing，自行運算。

用四則運算的結果則無法 missing，若欲 missing 則該比資料無法加總。

函數名稱	舉例
最小值	Y=MIN(of x1-x7);
最大值	Y=MAX(of x1-x7);
全距	Y=RANGE(of x1-x7);
總和	Y=SUM(of x1-x7);
平均值	Y=MEAN(of x1-x7);
標準差	Y=STD(of x1-x7);

變異數	Y=VAR(of x1-x7);
偏態係數	Y=SKEWNESS(of x1-x7);
峰度係數	Y=KURTOSIS(of x1-x7);

以上是算出每比資料均會有一個統計值，而非每個變項有一個統計值。

例如：需以父母最高教育程度代表家庭教育程度時，可寫

Fam_edu = **MAX** (**OF** mom_edu dad_edu);

3. 文字函數

用來處理文字變項的函數。

函數名稱	舉例	說明
壓縮文字間的空格	x='a b c d'; y=**COMPRESS**(x);	空格被壓縮掉了 Y='abcd'。
把文字置左	x=' HI'; y=**LEFT**(x);	把 X 內的字置左 Y='HI'。
把文字置右	x=' HI '; y=**RIGHT**(x);	把 X 內的字置右 Y='HI'。

抽取字串	x='19750913';
	year=**SUBSTR**(x, 1,4);
	month=**SUBSTR**(x, 5,2);
	day=**SUBSTR**(x, 7,2);
		年是從第 1 欄開始，連續讀取 4 欄，月是第 5 欄開始，連續讀取 2 欄，月是第 7 欄開始，連續讀取 2 欄。

合併字串	x='wu,'; z='wen-chi'; y=**TRIM**(x)||**TRIM**(z);	把 x 和 z 合併成 y=wu, wen-chi。
改變成數值格式	x='1234'; y= **INPUT**(x, 8.0);	把 x 變成數值 y。
改長度	**LENGTH** x $ 8.0; x='1234';	把 x 的長度設定成 8 位元 若不寫「$」，也會同時改成數值格式。 此指令可用來將變項排成所需要的順序。

（三）更改變項

1. 新變項和舊變項並存

   EX: 新變項=舊變項*2;

   新變項等於做一些運算後的舊變項。

2. 舊變項改名字

   RENAME 舊變項 1=新變項 1 舊變項 2=新變項 2;

六、 向量(ARRAY)運算

（一）ARRAY 的定義

1. 通式：ARRAY　名稱　{變項數目}　<$>　變項串;
2. 用意：用 ARRAY 可以一次做很多變項
3. 注意：用 ARRAY 會將原變項名稱蓋掉，保險起見，應另存新檔。
4. 程式：用 ARRAY 搭配 DO…END，或 IF…THEN，來改變一群變項。用「i」設定座標，通常等於要改變的變項之數目。
5. 附註：VAR1 至 VAR10，變項名稱有數字連續時，僅需以「-」相隔，（例：VAR1-VAR10）。若變項名稱完全不同，僅位置連在一起，則需以「--」相隔，若位置不相連，則需一個一個變項寫出來。\

（二）用 ARRAY 反向計分　[Prog06.sas]

用 ARRAY	一般寫法及備註

ARRAY q {*} q1-q5;
　　DO i=**1 to 5**;
　　　　q(i)=**6** - q(i);
　　END;
DROP i;
	q1=6-q1;
	q2=6-q2;
	q3=6-q3;
	q4=6-q4;
	q5=6-q5;

ARRAY q {*} q1-q10; 　　**DO** I=**1**,**3**,**5**,**7**,**9**; 　　　　q(i)=**6**-q(i); 　　**END**; **DROP** i;	跳著選變項 座標需要加上「,」

（三）用 ARRAY 產生新變項

用 ARRAY	備註
ARRAY sale {*} sale1-sale4; **ARRAY** change {*} change1-change3; 　　**DO** i=**1 to 3**; 　　　　change{i}=sale{i+**1**}-sale{i}; 　　**END**; **DROP** i;	新變項是 change1~ change3 舊變項是 sale1-sale4

（四）用 ARRAY 加總量表

用 ARRAY	備註

total=0;
ARRAY k{19} y7soc1--y7soc19;
DO i=1 TO 19;
IF k{i}>0 THEN total =total+k{i};
END;
DROP i;
	total 是總分
	將 19 題加總
	僅加有填答案的選項，missing 跳過

（五）用 ARRAY 計算平均值

用 ARRAY	備註

total=0;
n=0;
ARRAY k{19} y7soc1--y7soc19;
DO i=1 TO 19;
IF k{i}>0 THEN DO;
total=total+k{i};
n=n+1;
END;
END;
DROP i;
IF n>9 THEN mean=total/n;
	n 是題數
	最後一行表示僅答題數目超過一半者，才納入計算平均。

ARRAY test {*} test1-test6;
DO i=1 TO 6 BY 2;
right=MEAN(right, test(i));
left=MEAN(left, test(i+1));
END;
DROP i;
	共有六個變項。
	計算座標時每次跳兩格。
	積數欄位的變項平均值為 right，偶數欄位變項的平均值為 left。

（六）用 ARRAY 處理遺漏值

用 ARRAY	備註
ARRAY k{*} var1-var10; **DO** i=**1 TO 10;** **IF** k{i}=**99 THEN** k{i}=**.;** **END;** **DROP** i;	將所有原本等於 99 的值變成「.」
total=**0**; n=**0**; **ARRAY** k{**19**} y7soc1--y7soc19; **DO** i=**1 TO 19**; **IF** k{i}>**0 THEN DO**; total=total+k{i}; n=n+**1**; **END**; **END**; **DROP** i; **IF** n>**9 THEN** total_new=total/n***19**;	n 是題數 最後一行表示： (1)僅答題數目超過一半者，才納入計算平均，之後乘以題數，算新的總和。 (2)用自己有填答的部分取代自己的 missing。

參、SAS 的 PROC step 常用程式

一、 與結果有關的程式

（一）FORMAT 和 LABEL 和 TITLE　[Prog07.sas]

1. 用途：將變項給予文字說明，將選項給予文字說明。
2. 程式：

```
TITLE 'This is an illustration of PROC FORMAT and FORMAT
      statements';
PROC FORMAT ;
     VALUE $gender 'f'=女' 'm'='男';
     VALUE status 1='一年級' 2='二年級' 3='三年級' 4='四年級';

DATA roster;
     INFILE
     'E:\Wen_chi\CABLE\Cable2005EducationTrainning\Data\roster.dat';
     INPUT name $ sex $ id stand pretest first second final;
     FORMAT sex $gender.
            stand status.;
    LABEL sex = '性別'
          stand = '年級';
run;

PROC FREQ DATA=roster;
     TABLES sex*stand;
TITLE2 'Results of PROC FREQ';
RUN;
```

3. 程式註解：

(1)用 **PROC FORMAT** 配上 **FORMAT**，設定選項的文字說明。

(2)用 **LABEL** 設定變項的文字說明。

(3)在 PROC step 的裡面或外面，均可以用 **TITLE** 來註解，可以產生兩層標題的效果。

（二）PROC CONTENTS

1. 用途：瞭解資料的結構及內含變項。
2. 程式：

```
PROC CONTENTS DATA=roster VARNUM;
```

3. 程式註解： **VARNUM** 讓 PROC CONTENTS 的結果照變項在資料內的順序排列。

（三）PROC PRINT

1. 用途：將一些想要看的某部分資料印出來。
2. 程式：

```
PROC PRINT DATA=roster;
        VAR _all_;
        SUM pretest first second final;
        WHERE id<10
        run;
```

3. 程式註解：用 **VAR** 選要 PRINT 的變項，用 **SUM** 算一些數值變項的總和，用 **WHERE** 設定要列出的變項的條件。

二、 單變項分析

單變項和雙變項分析除瞭解變項分佈外，可用於除錯，用單變項看是否有不應該出現的值，用雙變項檢查有無跳答上的錯誤。

（一）連續變項 PROC MEANS 及 UNIVARIATE [Prog08.sas]

1.PROC MEANS：

(1)用來計算所有連續變項的相關屬性，例如平均值、標準差、變異數等。

程式：

```
PROC MEANS DATA=ACHIEVE MAXDEC=4
    N MIN MAX RANGE MEAN VAR STD STDERR
    SKEWNESS KURTOSIS;
RUN;

PROC SORT DATA=achieve; BY sex;
PROC MEANS DATA=ACHIEVE MAXDEC=4;
    VAR reading punc;
    CLASS grade;
    BY sex;
RUN;
```

註解：

MAXDEC：小數顯示至第四位，**N**：樣本數，**MIN**：最小值，**MAX**：最大值，**RANGE**：全距，**MEAN**：平均數，**VAR**：變異數，**STD**：標準差，**STDERR**：標準誤（平均數的標準差，為樣本標準差除以樣本數開根號），**SKEWNESS**：偏態係數（>0 表正偏或右偏，<0 表負偏或左偏），**KURTOSIS**：峰度（>0 比常態尖，<0 比常態平緩）。

前面五項為預設的結果。

搭配 **PROC SORT**，及 **CLASS** 和 **BY** 的選項，可做出分組的表，表示不同性別下，不同年級各做一次 PROC MEANS。

(2)計算獨立母體的 t-test，檢定某變項的平均數是否為 0。

程式：

```
PROC MEANS DATA=ACHIEVE MAXDEC=4 T PRT;
    VAR reading punc;
RUN;
```

註解：

T 表示要進行單一母體的 t-test，**PRT** 是列出 probability of t-test。

2.PROC UNIVARIATE

呈現所有連續變項有關的統計值及可粗略的繪圖。

```
PROC UNIVARIATE DATA=achieve ROUND=.01 PLOT
NORMAL;
    VAR reading;
run;
```

ROUND 表示結果呈現到小數兩位，**PLOT** 是畫圖（莖葉圖、盒狀圖及常態分佈圖），**NORMAL** 做是否符合常態分佈的檢定，大部分的檢定均是假設該變項屬常態分佈，若 p-value 有顯著，則推翻常態分佈。

（二）類別變項 PROC FREQ

1.一元次數分配表

```
PROC FREQ DATA=g;
    TABLES grade race course;
    WEIGHT freq;
RUN;
```

2.二元次數分配表

```
PROC FREQ DATA=g;
TABLES grade*(race course)/
NOCOL NOROW NOCUM NOFREQ NOPERCENT;
    WEIGHT freq;
```

```
RUN;
```

註解：二元列表，寫在前面的是列，後面是欄。可以在一直往上加，但在最後面的還是欄，往前加的都會變成分表。**NOCOL** 不顯示欄加總，**NOROW** 不顯示行加總，**NOCUM** 不顯示小計， **NOFREQ** 不顯示每個細格數，**NOPERCENT** 不顯示百分率。

三、 雙變項分析

（一）類別變項和類別變項的關係 [Prog09.sas]

1. 卡方檢定（Chi-square Test）

   檢定兩個類別變項間是否無關，若拒絕此假設，表示此兩類別變項有關。當 p<0.05 時（通常顯著水準是以 0.05 為標準，但可依研究者自己定義而改變。），表示 reject hypothesis，因此兩個變項有關。

2. 程式：PROC FREQ/ CHISQ

```
PROC FREQ DATA=chrt2k;
    TABLES area*y7cphone/EXPECTED CHISQ;
RUN;
```

   台北和新竹的國中生，擁有手機的比率是否有差異。

3. 程式註解：

   用 **PROC FREQ** 來做卡方檢定。

   EXPECTED 是讓 SAS 列出細格內的期望值，**CHISQ** 是進行卡方檢定。

4. 公式：

$$\chi^2 = \sum_{\text{所有細格}} \frac{(\text{期望次數} - \text{觀察次數})^2}{\text{觀察次數}}$$，自由度＝（欄數-1）×（列數-1）

   保守建議，若自由度為 1，則每個細格之期望次數至少要 10。若自由度大於或等於 2，則最低期望次數應為 5。

5. 結果摘錄：

area(地區)　　y7cphone(手機)

Frequency / Expected / Percent / Row Pct / Col Pct	沒有	有	Total
台北	466	614	1080
期望值 / 1031*52.89%	**545.29**	**534.71**	
	22.82	30.07	52.89
	43.15	56.85	
	45.20	60.73	
新竹	565	397	962
	485.71	**476.29**	
	27.67	19.44	47.11
	58.73	41.27	
	54.80	39.27	
Total	1031	1011	2042
	50.49	49.51	100.00

Frequency Missing = 130

Statistics for Table of area by y7cphone

Statistic	DF	Value	Prob
Chi-Square	1	49.4292	**<.0001**
Likelihood Ratio Chi-Square	1	49.6394	<.0001
Continuity Adj. Chi-Square	1	48.8078	<.0001
Mantel-Haenszel Chi-Square	1	49.4050	<.0001
Phi Coefficient		-0.1556	
Contingency Coefficient		0.1537	
Cramer's V		-0.1556	

卡方檢定值

Phi 係數，可看類別變項間關係的強度，若兩者無關則為 0。

Fisher's Exact Test

Cell (1,1) Frequency (F)	466
Left-sided Pr <= F	1.274E-12
Right-sided Pr >= F	1.0000
Table Probability (P)	6.029E-13
Two-sided Pr <= P	2.253E-12

Effective Sample Size = 2042
Frequency Missing = 130

若期望值太小時，SAS 會有 error 訊息，則需要看校正後的卡方檢定即 Fisher exact test。在 2x2 的表格時，SAS 會自動執行 Fisher exact test。在 option 處加上「**EXACT**」則會強制執行 Fisher exact test。

（二）兩類之類別變項與連續變項的關係 PROC TTEST

1. T 檢定：

（1） 可做兩組獨立樣本、相依樣本及一組樣本的 T 檢定。

（2） 假設：母體呈常態分佈、兩個獨立樣本來自變異數相等的母體。

2. 程式

(1)兩組獨立樣本

```
PROC TTEST DATA=chrt2k;
   CLASS area;
   VAR y7pmamnt;
run;
```

台北和新竹的國中生，每月零用錢的平均值有無差異。

(2)兩組相依樣本

```
PROC TTEST DATA=train.hight;
   PAIRED y7hight*y8hight;
run;
```

今年的身高和明年的身高之相關。

(3)單一樣本

```
PROC TTEST DATA=chrt2k;
   var y7hight;
run;
```

身高的平均值是否為零。（跟 PROC MEANS 的 T test 一樣）

3. 程式註解

用 **PROC TTEST** 做 T 檢定，用 **CLASS** 設定類別變項，用 **VAR** 設定連續變項，用 **PAIRED** 做相依樣本檢定。

4. 結果摘錄（兩組獨立樣本）

Variable	area	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err
y7pmamnt	台北	462	694.05	778.82	863.59	870.98	927.15	991.13	43.135
y7pmamnt	新竹	381	684.03	796.77	909.5	1044.9	1119.1	1204.8	57.335
y7pmamnt	Diff (1-2)		-156.3	-17.95	120.38	971.96	1018.4	1069.5	70.476

3.兩個地區的零用錢之平均值。

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
y7pmamnt	Pooled	Equal	841	-0.25	0.7991
y7pmamnt	Satterthwaite	Unequal	737	-0.25	0.8026

2.根據變異數結果選擇不同的 ttest 結果，此處應選 unequal 的。

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
y7pmamnt	Folded F	380	461	1.46	<.0001

1.先看變異數是否相等，若 reject 則表示變異數不相等。

（三）三類以上類別與連續變項的關係 PROC ANOVA 或 GLM

1. ANOVA 變異數分析（Analysis of Variance）

比較三個或三個以上母體平均數是否相等。

2. 程式

(1) 一因子變異數分析（One way ANOVA）

```
PROC ANOVA DATA=chrt2k;
   CLASS date;
   MODEL y7pmamnt=date;
   MEANS date/SCHEFFE;
RUN;
```

約會經驗與零用錢多少是否有關。

(2) 二因子變異數分析（Two way ANOVA）

```
PROC GLM DATA=chrt2k;
   CLASS date medu;
   MODEL y7pmamnt=date medu date*medu;
   MEANS date medu/SCHEFFE;
RUN;
```

看媽媽教育程度及約會經驗與零用錢多少是否有關。

3. 程式註解

此處用 **ANOVA** 或 **GLM** 均可，結果是一樣的，用 **CLASS** 設定類別變項，用 **MODEL** 寫一個 Y=X1 的模式，這裡的 Y 一定要是連續變項，X 一定要是類別變項。用 **MEANS** 做事後檢定，看看究竟是哪些組別

間有差異，**SCHEFFE** 是雪費事後檢定，還有其他的事後檢定，如 **LSD**、**TUKEY** 等，可以查 HELP。

4. 結果摘錄

The ANOVA Procedure

Dependent Variable: y7pmamnt　每月零用錢

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5710274.6	2855137.3	2.78	0.0625
Error	846	868440156.6	1026525.0		
Corrected Total	848	874150431.2			

R-Square	Coeff Var	Root MSE	y7pmamnt Mean
0.006532	129.3916	1013.176	783.0306

Source	DF	Anova SS	Mean Square	F Value	Pr > F
date	2	5710274.621	2855137.311	2.78	0.0625

Comparisons significant at the 0.05 level are indicated by ***.

date Comparison	Difference Between Means	Simultaneous 95% Confidence Limits	
最近一個月有 - 一個月前有	**159.30**	-466.64	785.24
最近一個月有 - 從來沒有	**391.23**	-134.43	916.88
一個月前有 - 最近一個月有	-159.30	-785.24	466.64
一個月前有 - 從來沒有	**231.93**	-130.56	594.42
從來沒有 - 最近一個月有	-391.23	-916.88	134.43
從來沒有 - 一個月前有	-231.93	-594.42	130.56

比較結果：最近一個月有的零用錢平均值 > 一個月前有的零用錢 > 從來沒有的零用錢

（四）連續變項與連續變項的關係 PROC CORR

1. 相關係數的分析（Correlation）

相關係數用來表示兩個變項間相關程度的強弱。

(1) 皮爾森簡單相關係數（Pearson correlation）：衡量 X 和 Y 數值相對其平均數，同時上升或下降的程度，表示 X 與 Y 間線性關係的強弱。通常用在連續變項。

(2) 斯皮爾曼等級相關係數 (Spearman correlation)：表示 X 和 Y 各自排序後，X 和 Y 的等級差異。可適用於序位變項。

(3) 信度--內部一致性（Cronbach's alpha）：用來量化一個測驗的信度，適用於檢測一個量表。

2. 程式

(1) 皮爾森簡單相關係數

```
PROC CORR DATA=chrt2k PEARSON;
  VAR y7hight y7weight;
RUN;
```

身高和體重的關係強弱。

(2) 斯皮爾曼等級相關係數

```
PROC CORR DATA=chrt2k SPEARMAN;
    VAR y7soc1--y7soc5;
RUN;
```

生活適應量表內，前五題的相關性。

(3) 信度--內部一致性

```
PROC CORR DATA=chrt2k ALPHA NOSIMPLE NOCORR;
    VAR y7soc1--y7soc19;
RUN;
```

3. 程式註解

用 **PROC CORR** 做所有有關相關性的分析，在 **OPTION** 的地方放置 **PEARSON**、**SPEARMAN**、**ALPHA** 可做出三種不同的分析。用 **NOSIMPLE** 去除描述性結果，用 **NOCORR** 去除相關係數的表。

4. 結果摘錄

(1) 皮爾森簡單相關係數

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	y7hight	y7weight	y7soc
y7hight 身高	1.00000 2100	0.58848 <.0001 2085	0.00227 0.9171 2100
y7weight 體重	0.58848 <.0001 2085	1.00000 2102	-0.03067 0.1599 2102
y7soc	0.00227 0.9171 2100	-0.03067 0.1599 2102	1.00000 2172

相關係數
P-value
樣本數

(2) 斯皮爾曼等級相關係數

Spearman Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	y7soc1	y7soc2	y7soc3	y7soc4	y7soc5
y7soc1 1做事帶來歡笑	1.00000 2171	0.42731 <.0001 2165	-0.01355 0.5298 2153	-0.28110 <.0001 2163	0.26751 <.0001 2160
y7soc2 2需要幫助時有人幫助	0.42731 <.0001 2165	1.00000 2165	0.04594 0.0332 2149	-0.21786 <.0001 2158	0.31437 <.0001 2155
y7soc3 3喜歡看電視	-0.01355 0.5298 2153	0.04594 0.0332 2149	1.00000 2153	0.05941 0.0059 2147	-0.00522 0.8093 2143
y7soc4 4做事覺得無聊	-0.28110 <.0001 2163	-0.21786 <.0001 2158	0.05941 0.0059 2147	1.00000 2163	-0.13505 <.0001 2153
y7soc5 5關心周圍的事	0.26751 <.0001 2160	0.31437 <.0001 2155	-0.00522 0.8093 2143	-0.13505 <.0001 2153	1.00000 2160

相關係數
P-value
樣本數

(3) 信度--內部一致性

Cronbach Coefficient Alpha

Variables	Alpha
Raw	0.505079
Standardized	0.519305

標準化分數後的內部一致性

Cronbach Coefficient Alpha with Deleted Variable

若去除此變項，內部一致性的改變。

Deleted Variable	Raw Variables: Correlation with Total	Raw Variables: Alpha	Standardized Variables: Correlation with Total	Standardized Variables: Alpha	Label
y7soc1	0.184852	0.487675	0.209830	0.497515	1做事帶來歡笑
y7soc2	0.180758	0.487920	0.212421	0.497033	2需要幫助時有人幫助
y7soc3	0.109262	0.500527	0.096752	0.518159	3喜歡看電視
y7soc4	0.054227	0.509231	0.018856	**0.531941**	4做事覺得無聊
y7soc5	0.251904	0.474140	0.276439	0.484993	5關心周圍的事
y7soc6	0.082284	0.505328	0.057025	**0.525232**	6朋友讓你失望
y7soc7	0.097337	0.504655	0.077392	**0.521618**	7沒有被公平對待
y7soc8	0.170611	0.489372	0.176038	0.503764	8喜歡冰淇淋
y7soc9	0.244391	0.475758	0.280591	0.484204	9事情困難但最後會解決
y7soc10	0.140111	0.494974	0.117201	0.514482	10感到困惑腦筋糾結
y7soc11	0.310315	0.463883	0.328858	0.474947	11瞭解朋友需要
y7soc12	0.246507	0.477097	0.275912	0.485093	12解決自己的問題
y7soc13	0.270349	0.470900	0.296583	0.481153	13對很多事有興趣
y7soc14	0.156855	0.491864	0.137109	0.510879	14力不從心
y7soc15	0.056266	0.512048	0.056980	**0.525240**	15喜歡醫護人員打針
y7soc16	0.175955	0.488077	0.199219	0.499485	16瞭解別人為什麼憤怒
y7soc17	0.098376	0.505533	0.077557	**0.521588**	17不滿意自己
y7soc18	0.098855	0.505589	0.088298	0.519672	18上課時不知道做什麼
y7soc19	0.166340	0.490231	0.178278	0.503352	19 想要一件東西確定得到

刪掉粗體的變項，改變沒有很大。

四、 迴歸分析

（一）依變項為連續變項之線性複迴歸分析 PROC REG

1. 線性複迴歸分析 Regression analysis

可同時看很多變項，連續和類別變項均可包括在內。

2. 公式

$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + ... + \beta_k x_{nk} + e_n \qquad e_n \sim N(0,\sigma^2)$$

$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 D1_{n2} + \beta_2 D2_{n2} + \beta_2 D3_{n2} + ... + \beta_k x_{nk} + e_n$$

若 X_{n2} 是類別變項，有四類時，必須製造三個 dummy variable 代替。

因為類別變項的數字沒有意義，dummy variable 的數目是類別變項的類數減一。。

迴歸係數的意義是在控制其他變項之下，改變該變項一單位時，依變項會增加或減少多少。

3. 假設：

(1) Y 在控制了所有的 X 後成常態分佈，即 error term 成常態。

(2) Y_n 之間是互相獨立的，即 e_n 互相獨立。

4. 程式

(1)一般複迴歸

```
PROC REG DATA=chrt2k;
   MODEL y7soc=y7kidsex area y7cphone shd1 shd2 dated1 dated2
             medud1 medud2/P R;
RUN;
PROC REG DATA=chrt2k /*(obs=20)*/;
   MODEL y7soc= y7hight y7weight/VIF STB;
RUN;
```

生活適應和性別、地區別、有無手機、自覺健康（好相對於不好、普通相對於不好）、約會經驗（最近一個月有相對於從來沒有、一個月前有過相對於從來沒有）的關係。

(2)逐步複迴歸

```
PROC REG DATA=chrt2k;
   MODEL y7soc=y7kidsex area y7cphone shd1 shd2 dated1 dated2
             medud1 medud2/
             SELECTION=STEPWISE sle=0.15 sls=0.05;
RUN;
```

5. 程式註解

用 PROC REG 作複迴歸分析，用 MODEL 建立一個 Y＝X1+X2+X3 之類的模式，用 P 列出預測值，用 R 列出殘差估計，VIF 檢測連續型自變項的共線性，超過 10 就是高共線性。用 SELECTION 選擇在模式中放置變項的方法，包括 FORWARD、BACKWARD、STEPWISE，SLE 設定排除的顯著水準，SLS 設定選入的顯著水準。

6. 結果

(1)一般複迴歸

The REG Procedure
Model: MODEL1
Dependent Variable: y7soc

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	1175.11899	130.56878	3.60	0.0002
Error	1788	64783	36.23230		
Corrected Total	1797	65958			

Root MSE	6.01933	R-Square	0.0178
Dependent Mean	59.69238	Adj R-Sq	0.0129
Coeff Var	10.08391		

橫斷>0.18
縱貫>0.6

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	57.73926	0.94559	61.06	<.0001
y7kidsex		1	0.97647	0.28950	3.37	**0.0008**
area	地區	1	-0.17560	0.30212	-0.58	0.5612
y7cphone	手機	1	0.40583	0.29277	1.39	0.1659
SHd1		1	-0.60865	0.49326	-1.23	0.2174
SHd2		1	0.55396	0.45547	1.22	0.2241
dated1		1	0.75402	0.62962	1.20	0.2312
dated2		1	0.86662	0.96721	0.90	0.3704
medud1		1	0.01766	0.39624	0.04	0.9644
medud2		1	-0.41344	0.47611	-0.87	0.3853

結果發現，只有性別和生活適應有關，女生(coding=2)比男生(coding=1)生活適應得分多 0.98 分。

The REG Procedure
Model: MODEL1
Dependent Variable: y7soc

Output Statistics

Obs	Dep Var y7soc	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2-1 0 1 2
1	54.0000	59.5176	0.3424	-5.5176	6.010	-0.918	| *| |
2	53.0000	58.3550	0.3992	-5.3550	6.006	-0.892	| *| |
3	57.0000	59.9234	0.3466	-2.9234	6.009	-0.486	| | |
4	49.0000	58.7608	0.4074	-9.7608	6.006	-1.625	| ***| |
5	59.0000	58.5325	0.5332	0.4675	5.996	0.0780	| | |
6	60.0000	58.9636	0.4994	1.0364	5.999	0.173	| | |
7	60.0000	58.9384	0.5375	1.0616	5.995	0.177	| | |
8	55.0000	.	.	.	.	.	| | |
9	61.2222	59.4923	0.3750	1.7299	6.008	0.288	| | |
10	54.0000	59.0865	0.3696	-5.0865	6.008	-0.847	| *| |
11	54.0000	59.4923	0.3750	-5.4923	6.008	-0.914	| *| |
12	62.0000	.	.	.	.	.	| | |
13	62.0000	58.9460	0.5976	3.0540	5.990	0.510	| |* |
14	60.0000	.	.	.	.	.	| | |
15	67.0000	59.3315	0.3975	7.6685	6.006	1.277	| |** |
16	68.0000	.	.	.	.	.	| | |
17	57.0000	60.8999	0.3174	-3.8999	6.011	-0.649	| *| |
18	50.0000	60.8999	0.3174	-10.8999	6.011	-1.813	| ***| |
19	56.0000	.	.	.	.	.	| | |
20	73.0000	60.8999	0.3174	12.1001	6.011	2.013	| |**** |

Obs	Cook's D
1	0.000
2	0.000
3	0.000
4	0.001
5	0.000
6	0.000
7	0.000
8	.
9	0.000
10	0.000
11	0.000
12	.
13	0.000
14	.
15	0.001
16	.
17	0.000
18	0.001
19	.
20	0.001

Cook's D 可檢定 influential case，對迴歸分析影響力過大的樣本，大於 1 算嚴重。

(2)逐步複迴歸 (僅摘錄最後一步)

The REG Procedure
Model: MODEL1
Dependent Variable: y7soc

Stepwise Selection: Step 3

Variable y7cphone Entered: R-Square = 0.0150 and C(p) = 3.2025

C(p)接近參數個數，且最小時，會停下來。

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	986.62077	328.87359	9.08	<.0001
Error	1794	64972	36.21619		
Corrected Total	1797	65958			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	56.93612	0.60650	319171	8812.92	<.0001
y7kidsex	0.97271	0.28732	415.07569	11.46	0.0007
y7cphone	0.47027	0.28719	97.11131	2.68	0.1017
SHd2	0.97930	0.28916	415.37805	11.47	0.0007

Bounds on condition number: 1.0239, 9.1471

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	y7kidsex			1	0.0070	0.0070	13.6559	12.69	0.0004
2	SHd2			2	0.0065	0.0135	3.8827	11.77	0.0006
3	y7cphone		手機	3	0.0015	0.0150	3.2025	2.68	0.1017

（二）依變項為類別變項之邏輯斯迴歸分析 PROC LOGISTIC

1. Logistic Regression

當依變項為兩類時，需用 Logistic Regression。

2. 公式

$$LOGIT(p) = \log[\frac{p}{1-p}] = \alpha + \beta' x$$

把迴歸係數取 exponential，即 exp(B)，會變成勝算比 OR(Odds Ratio)，可看出暴露在某 X 下，相對於沒有暴露於某 X 下，Y 發生的機率差多少倍。

3. 程式

```
PROC LOGISTIC DATA=chrt2k  DESCENDING;
  CLASS y7cphone y7kidsex area;
  MODEL y7cphone= y7kidsex area medud1 medud2;
RUN;
```

4. 程式註解

用 PROC LOGISTIC 做 logistic regression，用 descending 選擇依變項設定為「發生機率」的方向，預設是以數字大的作為參考組，用 CLASS 設定類別變項，用 MODEL 設定統計模式。其 option 和一般 REG 幾乎相同，所以亦可用 selection=stepwise 之類的 option。若要做交互作用時，則建議最好將所有類別變項自己設定 dummy，且可用 EXPB 列出 OR 值。

5. 結果

The LOGISTIC Procedure

Model Information

Data Set	WORK.CHRT2K	
Response Variable	y7cphone	手機
Number of Response Levels	2	
Number of Observations	1800	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Response Profile

Ordered Value	y7cphone	Total Frequency
1	有	890
2	沒有	910

依變項的設定

Probability modeled is y7cphone='有'.

NOTE: 372 observations were deleted due to missing values for the response or explanatory variables.

Class Level Information

Class	Value	Design Variables
y7kidsex	1	1
	2	-1
area	台北	1
	新竹	-1

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	2497.108	2420.487
SC	2502.603	2447.965
-2 Log L	2495.108	2410.487

越小越好

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq

顯著差異表示模式可成立

Likelihood Ratio	84.6207	4	<.0001
Score	83.2253	4	<.0001
Wald	80.4441	4	<.0001

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
y7kidsex	1	40.7620	<.0001
area	1	35.3457	<.0001
medud1	1	2.5170	0.1126
medud2	1	1.2830	0.2573

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.2216	0.1202	3.3999	0.0652
y7kidsex	1	1	-0.3102	0.0486	40.7620	<.0001
area	台北	1	0.3018	0.0508	35.3457	<.0001
medud1		1	0.2143	0.1351	2.5170	0.1126
medud2		1	0.1835	0.1620	1.2830	0.2573

跟一般迴歸一樣看法

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
y7kidsex 1 vs 2		0.538	0.444	0.651
area	台北 vs 新竹	1.829	1.499	2.232
medud1		1.239	0.951	1.614
medud2		1.201	0.875	1.651

OR 在這裡！<1 是保護因子，>1 是危險因子，=1 是無關。CI 也可看出端倪，包括 0 則為無顯著相關。

性別與地區均和是否有手機有關，男生(coding=1)擁有手機的機率是女生(coding=2)的0.538倍，台北(coding=1)擁有手機的機率是新竹(coding=2)的1.829倍。

Association of Predicted Probabilities and Observed Responses

Percent Concordant	56.0	Somers' D	0.231
Percent Discordant	32.9	Gamma	0.260
Percent Tied	11.1	Tau-a	0.116
Pairs	809900	c	0.616

模式適配度，看預測值及觀測值比較後，準確的程度，C 值越高越好。

五、 多變量分析

（一）因素分析 PROC FACTOR

1. 因素分析 Factor analysis

   可自某幾個變項或題目中，萃取出幾個潛在的因素。這些題目或變項是可觀察的，潛在因素是運算出來的，不是直接觀察到的。

2. 程式

```
PROC FACTOR DATA=chrt2k ROTATE=varimax REORDER OUT=test2
            NFACTORS=4;
VAR y7soc1--y7soc19;
RUN;
```

3. 程式註解

用 ROTATE 轉軸，REORDER 將 factor loading 照大小排序，OUT 把因素分數存檔，NFACTORS 強制設定要分幾類因素。

4. 結果

Eigenvalues of the Correlation Matrix: Total = 19 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	3.71132254	1.80899455	0.1953	0.1953
2	1.90232799	0.55276884	0.1001	0.2955
3	1.34955915	0.29109409	0.0710	0.3665
4	1.05846507	0.06727371	0.0557	0.4222
5	0.99119136	0.04935074	0.0522	0.4744
6	0.94184062	0.06635302	0.0496	0.5239
7	0.87548760	0.01292682	0.0461	0.5700
8	0.86256078	0.05427050	0.0454	0.6154
9	0.80829028	0.02502609	0.0425	0.6579
10	0.78326419	0.01917512	0.0412	0.6992
11	0.76408906	0.06823482	0.0402	0.7394
12	0.69585424	0.01328753	0.0366	0.7760
13	0.68256671	0.00892322	0.0359	0.8119
14	0.67364349	0.02462696	0.0355	0.8474
15	0.64901654	0.01755302	0.0342	0.8816
16	0.63146352	0.04391102	0.0332	0.9148
17	0.58755250	0.06129420	0.0309	0.9457
18	0.52625831	0.02101226	0.0277	0.9734
19	0.50524605		0.0266	1.0000

4 factors will be retained by the MINEIGEN criterion.

現在的 eigenvalue 設定為 1，所以只要對於總變異量貢獻超過 1 以上的潛在因素都會被萃取出來。總解釋變異量 Cumulative 達 42.22%。

Rotated Factor Pattern

		Factor1	Factor2	Factor3	Factor4
y7soc11	11瞭解朋友需要	0.65889	0.09880	-0.02382	-0.05342
y7soc9	9事情困難但最後會解決	0.65424	-0.19741	0.01919	0.05204
y7soc5	5關心周圍的事	0.64248	0.00842	0.02729	-0.15728
y7soc12	12解決自己的問題	0.62859	-0.14194	-0.03133	0.08628
y7soc13	13對很多事有興趣	0.56022	-0.13935	0.16936	0.17778
y7soc16	16瞭解別人為什麼憤怒	0.55387	0.01600	-0.21787	-0.08123
y7soc2	2需要幫助時有人幫助	0.54534	-0.35272	0.18811	0.03008
y7soc1	1做事帶來歡笑	0.51354	-0.36646	0.11672	0.23714
y7soc17	17不滿意自己	-0.06331	0.65215	0.01877	-0.06900
y7soc10	10感到困惑腦筋糾結	-0.01733	0.60699	0.11161	-0.01886
y7soc4	4做事覺得無聊	-0.17769	0.59336	0.09972	-0.03269
y7soc7	7沒有被公平對待	-0.06446	0.58023	-0.09174	0.08102
y7soc6	6朋友讓你失望	-0.06719	0.56297	-0.06375	0.00726
y7soc14	14力不從心	-0.04658	0.48351	0.11898	0.32264
y7soc3	3喜歡看電視	-0.04915	0.07098	0.78265	-0.03045
y7soc8	8喜歡冰淇淋	0.10806	0.01670	0.73868	0.00209
y7soc15	15喜歡醫護人員打針	0.00561	-0.02963	-0.18459	0.72871
y7soc18	18上課時不知道做什麼	-0.12464	0.31007	0.20221	0.44742
y7soc19	19 想要一件東西確定得到	0.30380	-0.06679	0.03747	0.34638

通常以因素負荷量大於 0.3 為主，選擇在所有潛在因素中因素負荷量最大的那個因素作歸類，目前分出四類。

（二）集群分析 PROC FASTCLUS

1. 集群分析 Cluster analysis 是將人依據一些變項進行分類。

2. 程式

```
PROC FASTCLUS DATA=test2 OUT=test3 MAXC=2;
```

```
  VAR factor1--factor4;
RUN;
```

3. 程式註解

用 OUT 將 PROC CLUSTER 的結果存出來，檔名叫 test3，用 MAXC 設定要跑出幾個 cluster

4. 結果

Cubic Clustering Criterion = -12.302

CCC 值越小越好，可以試分成幾類 cluster 的 CCC 最小。

The FASTCLUS Procedure
Replace=FULL Radius=0 Maxclusters=2 Maxiter=1

Cluster Means

Cluster	Factor1	Factor2	Factor3	Factor4	
1	-.5927349626	-.3991348246	-.3445502581	-.3206292576	低
2	0.4669711434	0.3144482056	0.2714451451	0.2525995941	高

Cluster Standard Deviations

Cluster	Factor1	Factor2	Factor3	Factor4
1	0.906474956	0.872855936	0.995429719	0.824527798
2	0.803764996	0.981765542	0.916907976	1.052679045