報告題名：

套裝軟體 R 與 SAS 之優劣分析比較

作者：林貞妤、葉素婷、李玲佑、張惠雯、賀宇涵

系級：統計四乙

學號： D9639335、D9639127、D9660109、D9660010、D9645452

開課老師：陳婉淑 教授

課程名稱：統計專題

開課系所：統計系

開課學年：九十九 學年度 第 一 學期

中文摘要

有關統計最常使用的兩種軟體，一種是需要專業語法的 SAS，另一種則是隨處可得的免費軟體 R。為了找出對統計最有效率及分析良好的軟體，我們分別用 SAS 跟 R 對相同資料做出統計分析，並比較輸出結果，從中了解 SAS 跟 R 的優缺點，建議使用軟體，進而做為日後的參考。

利用某特定資料分別以 SAS 與 R 進行基本統計量、相關係數分析、散佈圖、多重共線性、離群值、選取重要變數、部分 F 檢定、偵測影響點、殘差診斷分析等等，配適複迴歸線所需要的檢定與統計分析，從中比較以 SAS 及 R 進行分析所需的程式碼及得到的報表有何不同，並且評斷 SAS 及 R 的差異性及優劣性。

從這份報告中，在 SAS 與 R 兩種統計套裝軟體都可以得到的情況下，可依不同的研究性質及方法，交互使用兩個套裝軟體 R 及 SAS，如針對基本統計量、相關係數分析、選取重要變數、殘差診斷分析可建議選用 SAS 程式；而散佈圖、多重共線性、離群值、部分 F 檢定、偵測影響點則建議選用 R 程式，以達到最有效率、省時且方便的統計分析。

關鍵字： **R、SAS、複迴歸、統計軟體**

目　　錄

表目錄

圖目錄

第一章　緒論

第一節　研究背景

統計學是一種工具，用來幫助人們蒐集、整理、分析資料，用以解釋及預測經濟、社會及自然現象的一門重要學問，已被廣泛的運用在保險業、企業、工業、醫學業、教育業，甚至是生態環境上，在各個科學中都扮演著不可或缺的重要角色。統計學為當今國外的熱門研究領域，國內也正緩慢起步當中。在現今科技資訊爆發的時代，透過統計學，不但可以提高研究與決策的品質，做出最合理、最科學的推論，更可以進一步提高生產力。統計學對我們日常生活幫助如此的廣泛，其發展不可限量。統計學是利用資料的特質來觀察並且了解一些現象，我們常藉由執行統計套裝軟體，以獲得統計分析結果。

經常使用的統計套裝軟體應該是 **SAS** 系統，最早由北卡羅來納州立大學的兩位生物統計學的研究生編製。起初 SAS 只是一數學統計軟體，後於 1976 年由 Jim Goodnight 博士及 John Sall 博士等人成立 SAS 公司，並正式推出 SAS 軟體。SAS 系統是一個模組軟體系統，它由多個功能的模組組合而成，最常使用的模組是 BASE 和 STAT。經過多年的發展，SAS 已經遍佈全世界，使用的單位遍及金融、醫藥衛生、生產、運輸、通訊、科學研究、政府和教育等領域；在資料處理和統計分析領域，SAS 系統被譽統計軟體界的巨無霸。

R 為免費的套裝軟體，且取得便利，本來是由來自紐西蘭奧克蘭大學的 Ross Ihaka 和 Robert Gentleman 開發，現在則由「R 開發核心團隊」負責開發。R 的語法是來自 Scheme，是基於 S 語言的一個 GNU 項目，所以也可以當作 S 語言的一種實現，通常用 S 語言編寫的代碼都可以不作修改在 R 環境下運行。**R 語言**主要運用於統計分析或者開發統計相關的軟體，不但如此也有人將 R 用作矩陣的計算。操作環境與繪圖功能亦是 R 強項，其製圖具有印刷的質素，也可在其中加入數學符號。R 內建多種統計學及數字分析功能，可以透過用戶撰寫的功能安裝套件（Packages）增強。由於 S 語言的血緣，R 比其他統計學或數學專用的編程語言，有更強的物件導向（物件導向程序設計）功能。

人們利用 R 及 SAS 進行統計分析，得到結果並做出適當的決策和解決之道，因此這兩個常用的統計套裝軟體 R 與 SAS 成為重要的學習課程。

第二節　研究動機

統計學經常被用在保險業、企業界、工業、醫學，甚至是教育上，廣泛的存在各個領域中，在日常生活中成為不可或缺的重要角色。因此統計套裝軟體的應用也成為重要的學習課程，我們常透過統計套裝軟體對感興趣的資料進行統計的

相關分析，但市面上的統計套裝軟體有很多種，卻不知該選擇何種統計套裝軟體來進行統計分析。常用的統計套裝軟體有 R、SAS，不同的軟體，跑出的報表數值或者是圖表也不盡相同，因此在進行統計資料分析時，常常思考著該使用哪種軟體進行分析，可以較簡單的得到較完整的結果，且有較美觀的圖表，已供進行分析並做出結論。

第三節　研究目的

統計套章軟體中常使用的有 R 與 SAS。SAS 沒有單機版，只有機構版，而且每年都必須計費且收費不便宜，換句話說，SAS 是提供給機構或團體使用，若不是在機構工作的研究人員或沒有與 SAS 簽約的機構研究人員都使用 SAS 上會很困難。R 不需花費就能得到的統計套裝軟體，可以自由取得。我們選擇對需要購買版權的 SAS 及隨處可得的免費軟體 R 進行討論研究。利用由心理與教育統計學中蒐集的一筆數據，分別使用 R 與 SAS 進行複迴歸分析，並且對於兩者報表進行優缺點的探討，並找出何種分析方法較適合使用何種統計套裝軟體，並且分析何種套裝軟體可以得到較美觀的圖，從中討論出 SAS 與 R 的優缺點，進而做為日後統計分析參考用。

第四節　研究方法

為了進行透過 SAS 和 R 做統計分析的優劣之比較，我們利用由心理與教育統計學中蒐集的一筆數據，分別使用 R 與 SAS 進行複迴歸線的配適，透過配適迴歸線的過程分析並且比較 R 與 SAS 的優缺點。

資料是由心理與教育統計學(林清山民 78)中，所找出的一筆資料數據，統計了 16 位學生的英文閱讀測驗成績，反應變數 Y 是學生的閱讀測驗成績，五個解釋變數分別為測驗中的單字成績(x1)、測驗中的片語成績(x2)、測驗中的文法成績(x3)、是學生對本次閱讀測驗的期望成績(x4)以及學生的智力測驗成績(x5)。

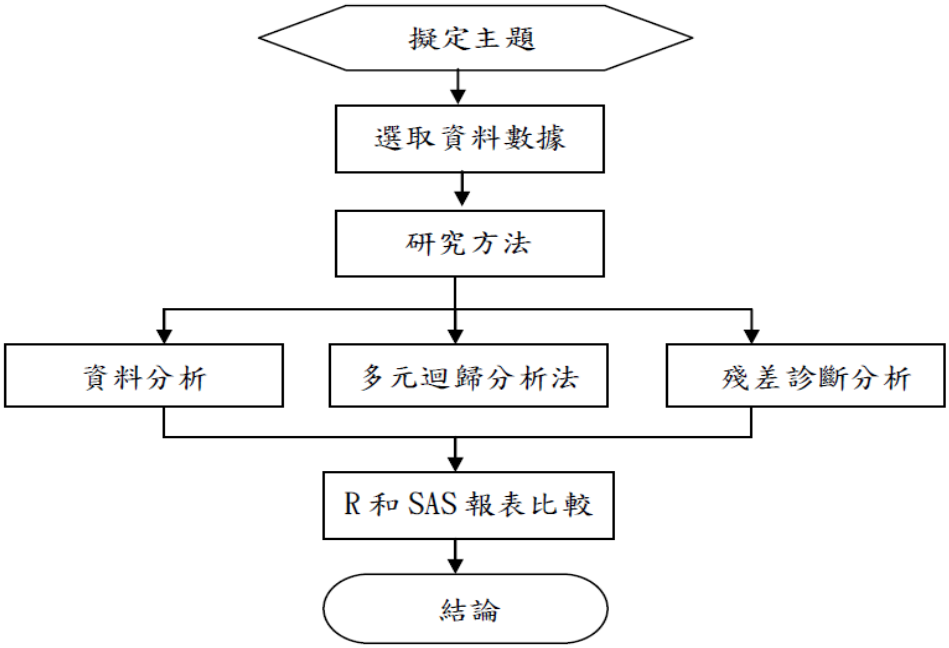
藉由此筆數據，我們透過 R 與 SAS 兩種統計套裝軟體進行複迴歸的統計分析。首先利用 R 與 SAS 進行資料分析，分別找出此筆資料的基本統計量、相關係數、散佈圖、離群值，並且偵測解釋變數間是否存在高度相關性。

接著進行迴歸線的配適，藉由向前選取法(Forward Selection)、向後選取法(Backward Selection)、逐步迴歸法(Stepwise regression methods)及全部子集迴歸法(All-subsets Regression)找尋資料的重要變數，選取重要變數後進行部分 F 檢定，檢定重要變數的選擇是否適當，接著找尋影響點。

最後對配適的迴歸線進行殘差檢定，分別有：殘差平均是否為零、殘差變異數是否為常數、檢定殘差是否相互獨立，以及檢定誤差是否為常態。檢定與假設相符進而確定最終模型。利用統計套裝軟體 R 與 SAS，進行以上的資料分析及複

迴歸線的配適，透過這些分析，比較兩統計套裝軟體需要的程式碼及得到的報表之差異與優劣。

第五節　研究流程



為了比較 R 和 SAS 統計套裝軟體，我們透過網路來選取資料，找到一筆有關於心理與教育統計學的數據，分別利用統計套裝軟體 R 和 SAS，進行統計分析，首先利用研究方法中的資料分析，來了解資料狀況，接著利用多元迴歸分析法配飾迴歸模型，最後對配適的模型進行殘差診斷分析，透過統計套裝軟體 R 和 SAS 比較兩統計套裝軟體需要的程式碼及得到的報表之差異與優劣，最後進行討論並下結論。

第二章　資料分析

第一節　資料的讀取

一、資料簡介

資料是由心理與教育統計學(林清山民 78)中，所找出的一筆資料數據，統計了 16 位學生的英文閱讀測驗成績，主要探討學生對單字、片語、文法的熟悉度以及智力的高低和預期自己的能力是否會和測驗出的成績高低有相關。反應變數 Y 是學生的閱讀測驗成績，五個解釋變數分別為測驗中的單字成績(x1)、測驗中的片語成績(x2)、測驗中的文法成績(x3)、是學生對本次閱讀測驗的期望成績(x4)以及學生的智力測驗成績(x5)。相關網址如下：

http://webclass.ncu.edu.tw/~tang0/Chap12/Sas12.htm#範例 12.5

二、SAS 與 R 的資料讀取之程式碼

1. SAS 程式碼：

```
data grade;
input y x1 x2 x3 x4 x5 ;
label y='閱讀測驗成績'
x1='單字成績'
x2='片語成績'
x3='文法成績'
x4='對閱讀的期望成績'
x5='智力測驗成績'
;
cards;
70 16 19 29 88 108
63 10 30 23 71 113
.
.
.
62 12 50 33 80 105
;
```

2. R 程式碼：

```
data=read.table(file='C:/score.txt',header=T)
## "header=T" 是以讀到的第一列做為變數名稱 ##
## 以下是對數據的命名 ##
y=data[,1]                    ## 閱讀測驗成績 ##
x1=data[,2]                   ## 單字成績 ##
x2=data[,3]                   ## 片語成績  ##
x3=data[,4]                   ## 文法成績 ##
x4=data[,5]                   ## 對閱讀測驗的期望成績 ##
x5=data[,6]                   ## 智力測驗成績 ##
```

第二節　一般基本統計量

一、何謂基本統計量

統計量是用來描述樣本特性的統計測量數，常被用來推估母體參數。一般統計量包括最基本的平均數、中位數、全距、變異數、標準差、偏態係數、峰態係數以及變異係數……等等，藉由這些基本統計量可以獲得資料的基本訊息。

1. 集中趨勢量數(平均數、中位數)

(1) 平均數：

一組數值加總後再除以總個數所得的值。

(2)中位數：

將資料值依大小順序排列，取其正中央之數值或正中央之兩數值之平均數。

2. 離散趨勢量數(全距、變異數、標準差、變異係數、偏態與峰態)

(1)全距(Range)：

樣本或母體中最大值與最小值的差，全距越大表資料的離散程度越大。

(2)變異數：

每個觀察值減去母體平均數(即離均差)，加以平方後加總，最後除以個數

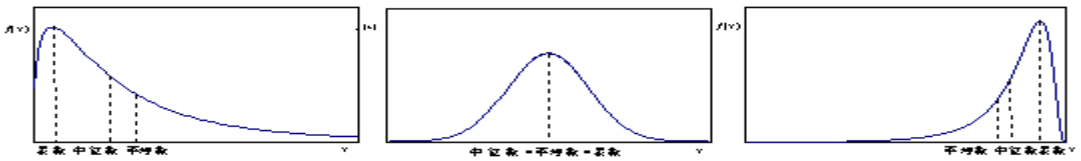
(Var(x)＝$\sum_{i=1}^{N}(x_i-\mu)^2/N$)，變異數越大表資料離散度越大。。

(3)標準差：

標準差為變異數之平方根。

(4)偏態係數：

偏態係數＞0，表右偏分配；偏態係數＝0，表對稱分配；偏態係數＜0，表左偏分配。其中，偏態係數＝$E[(x_i-\mu)^3]/\sigma^3$。



右偏分配　　對稱分配　　左偏分配

峰態係數＞3，表高峽峰；峰態係數＝3，表常態峰；峰態係數＜3，表低闊峰。其中，峰態係數＝$E[(x_i-\mu)^4]/\sigma^4$。

(6)變異係數：

標準差除以平均數就是變異係數用以比較兩組資料相對離散程度的工具。

變異係數＝$\sigma_x/\mu\times 100\%$

二、SAS 與 R 的基本統計量之程式碼

1. SAS 程式碼：

```
proc means data=grade RANGE Q1 Q3 MAX MIN MEAN MEDIAN CV STD KURT SKEW;
run;
```

2. R 程式碼：

```
summary(data)          ##   min Q1 median Q3 max   ##
mean(data)             ## 平均 ##
sd(data)               ## 標準差 ##
library(moments)       ## 峰態、偏態需用到此程式套件 ##
kurtosis(data)         ## 峰態 ##
skewness(data)         ## 偏態 ##
```

三、SAS 與 R 報表的分析及比較

1. 報表分析：

The MEANS Procedure

Variable	Label	Range	Lower Quartile	Upper Quartile	Maximum	Minimum
Y	閱讀測驗成績	37.0000000	59.5000000	72.5000000	81.0000000	44.0000000
x1	單字成績	14.0000000	11.5000000	19.5000000	22.0000000	8.0000000
x2	片語成績	59.0000000	30.0000000	45.5000000	78.0000000	19.0000000
x3	文法成績	30.0000000	22.5000000	33.0000000	40.0000000	10.0000000
x4	對閱讀的期望成績	49.0000000	71.0000000	84.5000000	92.0000000	43.0000000
x5	智力測驗成績	33.0000000	104.0000000	115.0000000	125.0000000	92.0000000

Variable	Label	Mean	Median	Coeff of Variation	Std Dev	Kurtosis
Y	閱讀測驗成績	65.2500000	66.0000000	16.8396507	10.9878721	-0.6004020
x1	單字成績	15.5000000	16.5000000	29.6118959	4.5898439	-1.3082223
x2	片語成績	38.3750000	37.0000000	37.9404493	14.5596474	2.6127019
x3	文法成績	27.7500000	28.0000000	30.2788491	8.4023806	-0.0803533
x4	對閱讀的期望成績	74.7500000	77.0000000	17.5959150	13.1529464	1.4704944
x5	智力測驗成績	109.8750000	110.0000000	8.4719892	9.3085982	-0.1899752

Variable	Label	Skewness
Y	閱讀測驗成績	-0.3276253
x1	單字成績	-0.1985672
x2	片語成績	1.1583574
x3	文法成績	-0.3693236
x4	對閱讀的期望成績	-1.2454375
x5	智力測驗成績	-0.1168057

表 2-2.1　基本統計量 SAS output

```
> summary(data)     ##  min Q1 median Q3 max  ##
       y                x1              x2              x3              x4              x5
 Min.   :44.00   Min.   : 8.00   Min.   :19.00   Min.   :10.00   Min.   :43.00   Min.   : 92.0
 1st Qu.:59.75   1st Qu.:11.75   1st Qu.:30.00   1st Qu.:22.75   1st Qu.:71.00   1st Qu.:104.5
 Median :66.00   Median :16.50   Median :37.00   Median :28.00   Median :77.00   Median :110.0
 Mean   :65.25   Mean   :15.50   Mean   :38.38   Mean   :27.75   Mean   :74.75   Mean   :109.9
 3rd Qu.:71.75   3rd Qu.:19.25   3rd Qu.:44.25   3rd Qu.:33.00   3rd Qu.:84.25   3rd Qu.:114.5
 Max.   :81.00   Max.   :22.00   Max.   :78.00   Max.   :40.00   Max.   :92.00   Max.   :125.0
> mean(data)            ## 平均 ##
      y      x1      x2      x3      x4      x5
 65.250  15.500  38.375  27.750  74.750 109.875
> sd(data)            ## 標準差 ##
        y        x1        x2        x3        x4        x5
10.987872  4.589844 14.559647  8.402381 13.152946  9.308598
> library(moments)       ## 峰態、偏態需用到此程式套件 ##
> kurtosis(data)      ## 峰態 ##
       y       x1       x2       x3       x4       x5
2.218537 1.713347 4.511811 2.589709 3.696588 2.511469
> skewness(data)      ## 偏態 ##
         y         x1         x2         x3         x4         x5
-0.2960737 -0.1794444  1.0468030 -0.3337563 -1.1254970 -0.1055568
```

表 2-2.2　基本統計量 R output

由表 2-1.1 SAS 的報表皆可得到各個變數的平均數、中位數、全距、變異數、標準差、變異係數、偏態係數與峰態係數。以解釋變數 x5 為例，其平均數為 109.875、中位數為 110、全距為 33、標準差為 9.3085、變異係數為 8.4719、偏態係數為－0.1168 及峰態係數為－0.1899。

表 2-1.2 R 的報表結果，除了峰態係數以外基本上都和 SAS 的結果相同，我們進一步的加以探討兩者的差異後發現，SAS 給訂的峰態係數是超額峰態(峰態係數－3)，而非一般的峰態係數，在此須特別注意。

2. 報表主觀的比較：

SAS 的 output 簡單清楚，變數間基本統計量的比較相對於 R 來的容易比對。因此在基本統計量上，我們喜愛使用 SAS 來進行分析，但要對於峰態係數要特別的注意，SAS 內定的是屬超額峰態。

第三節　相關係數

一、何謂相關係數

設有兩組樣本$x_1, x_2, \cdots, x_n$及$y_1, y_2, \cdots, y_n$，其樣本平均數分別為$\bar{x}$與$\bar{y}$，樣本標準差分別為S_x與S_y，且兩組樣本之樣本共變異數(covariance) S_{xy}，將其定義為 $S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$。則相關係數 r 定義為 $r = S_{xy}/S_x S_y = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$。透過相關係數分析可以得到兩變數間的相關性，並且可得到兩變數相關性的強度，相關係數絕對值越接近 1，表示兩變數的相關性越高。

二、SAS 與 R 的相關係數之程式碼

1. SAS 程式碼：

```
proc corr;
var v x1 - x5;
run;
```

2. R 程式碼：

```
cor(cbind(y,x1,x2,x3,x4,x5) )
```

三、SAS 與 R 報表的分析及比較

1. 報表分析：

The CORR Procedure

6 Variables: Y x1 x2 x3 x4 x5

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Y	16	65.25000	10.98787	1044	44.00000	81.00000	閱讀測驗成績
x1	16	15.50000	4.58984	248.00000	8.00000	22.00000	單字成績
x2	16	38.37500	14.55965	614.00000	19.00000	78.00000	片語成績
x3	16	27.75000	8.40238	444.00000	10.00000	40.00000	文法成績
x4	16	74.75000	13.15295	1196	43.00000	92.00000	對閱讀的期望成績
x5	16	109.87500	9.30860	1758	92.00000	125.00000	智力測驗成績

Pearson Correlation Coefficients, N = 16
Prob > |r| under H0: Rho=0

	Y	x1	x2	x3	x4	x5
Y 閱讀測驗成績	1.00000	0.28950 0.2768	-0.14523 0.5915	0.27223 0.3077	0.82524 <.0001	0.76358 0.0006
x1 單字成績	0.28950 0.2768	1.00000	-0.08180 0.7633	0.02247 0.9342	0.17890 0.5074	0.06710 0.8050
x2 片語成績	-0.14523 0.5915	-0.08180 0.7633	1.00000	0.02262 0.9337	-0.20731 0.4411	-0.38528 0.1406
x3 文法成績	0.27223 0.3077	0.02247 0.9342	0.02262 0.9337	1.00000	0.22199 0.4086	0.19562 0.4678
x4 對閱讀的期望成績	0.82524 <.0001	0.17890 0.5074	-0.20731 0.4411	0.22199 0.4086	1.00000	0.66892 0.0046
x5 智力測驗成績	0.76358 0.0006	0.06710 0.8050	-0.38528 0.1406	0.19562 0.4678	0.66892 0.0046	1.00000

表 2-3.1　相關係數 SAS output

```
> cor(cbind(y,x1,x2,x3,x4,x5) )
           y          x1          x2         x3          x4          x5
y  1.0000000  0.28949519 -0.14522690 0.27222885  0.8252441  0.76357756
x1 0.2894952  1.00000000 -0.08180386 0.02247252  0.1788969  0.06709575
x2 -0.1452269 -0.08180386  1.00000000 0.02261537 -0.2073083 -0.38527782
x3 0.2722288  0.02247252  0.02261537 1.00000000  0.2219888  0.19561619
x4 0.8252441  0.17889689 -0.20730826 0.22198885  1.0000000  0.66892367
x5 0.7635776  0.06709575 -0.38527782 0.19561619  0.6689237  1.00000000
```

表 2-3.2　相關係數 R output

由表 2-2.1 SAS 報表可知變數間是否存在相關性，以解釋變數 x4 與 x5 為例：在虛無假設$H_0: R = 0$，顯著水準為 0.05 之檢定下，p-value＝0.0046 < 0.05，拒絕H_0，表示兩變數間具有相關性。且由報表亦可知兩者的相關係數為 0.6689，成些微的正相關。

由表 2-2.2 R 的報表亦可得到與 SAS 報表相同的相關係數值，但 R 並不會自己進一步做相關係數是否為零的假設檢定，可能須找尋另一個程式碼才可得到此檢定結果。

2. 報表主觀的比較：

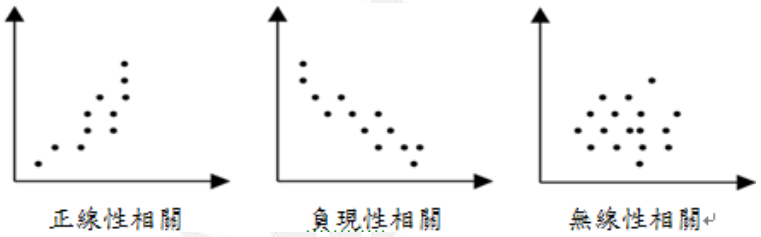
SAS 在一個指令下，不但可以跑出兩變數的相關係數，還會對兩變數是否存在相關性加以檢定($H_0: R = 0$)。而 R 則只是給定兩變數的相關係數，但單從相

關係數值來判斷是否存在相關性，並不是那麼的方便與準確，因此必須對此檢定輸入另一個指令，找到的結果加以判斷。希望得到相同的東西，SAS 只需一個指令即可，R 則非如此，因此在相關係數上，我們建議使用 SAS 進行分析。

第四節　散佈圖

一、何謂散佈圖

透過散佈圖不但可了解兩變數之間的相關性，亦可知道此筆資料是否存在異常值，它有直觀簡便的優點。但兩變數相關不代表就有因果關係，通過觀察相關圖主要是看點的分佈狀態，概略地估計兩因素之間有無相關關係，從而得到兩個變數的基本關係。當散佈圖的縱軸隨著橫軸的增加而增加，表示兩變數存在正線性相關；當縱軸隨著橫軸的增加而減少，表示兩變數存在負線性相關；而當橫軸的增加不會造成縱軸的增加或減少之趨勢時，表示兩變數並無線性相關。如下圖示：



二、SAS 與 R 散佈圖繪製的程式碼

1. SAS 程式碼：

```
proc gplot;
plot y*(x1 x2 x3 x4 x5);
symbol1 v=dot c=blue;
run;
```

2. R 程式碼：

```
win.graph()                    ## 此指令可使跑出的圖不重疊覆蓋出現 ##
par(mfrow=c(2,2))              ## 此指令可將畫面分割成 2*2 的四張圖 ##
plot(x1,y,main="scatterplot of (x1,y)")
plot(x2,y,main="scatterplot of (x2,y)")
plot(x3,y,main="scatterplot of (x3,y)")
plot(x4,y,main="scatterplot of (x4,y)")
win.graph()
plot(x5,y,main="scatterplot of (x5,y)")
```

三、SAS 與 R 報表的分析及比較

1. 報表分析：

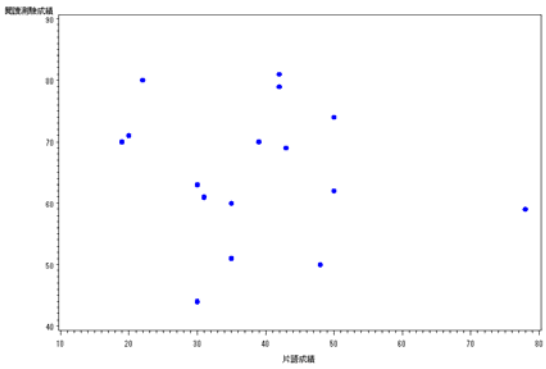


圖 2-4.1 x1 與 y 的散佈圖

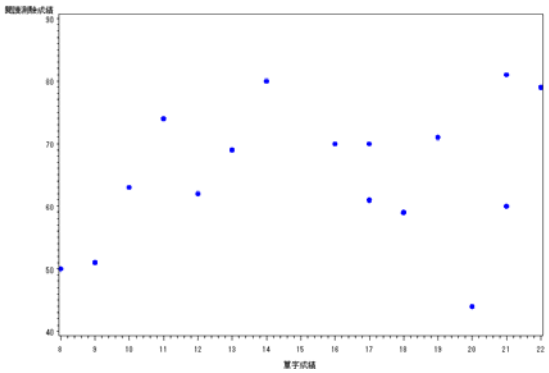


圖 2-4.2 x2 與 y 的散佈圖

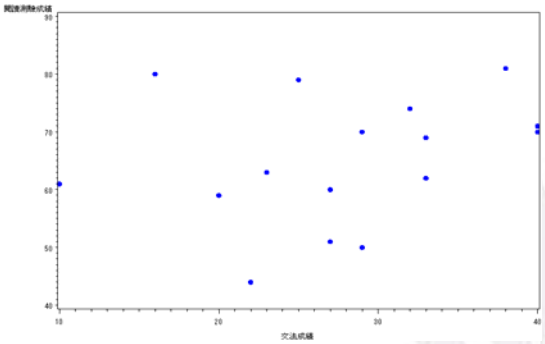


圖 2-4.3 x3 與 y 的散佈圖

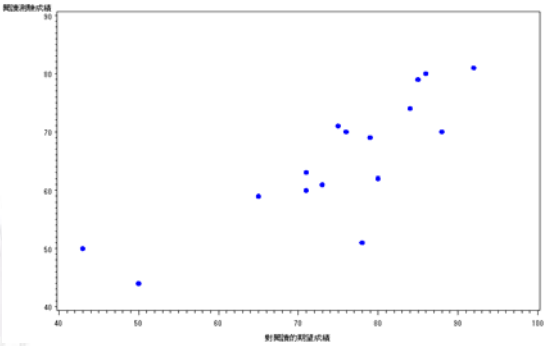


圖 2-4.4 x4 與 y 的散佈圖

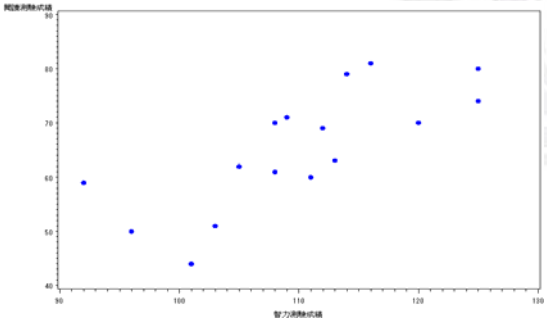


圖 2-4.5 x5 與 y 的散佈圖

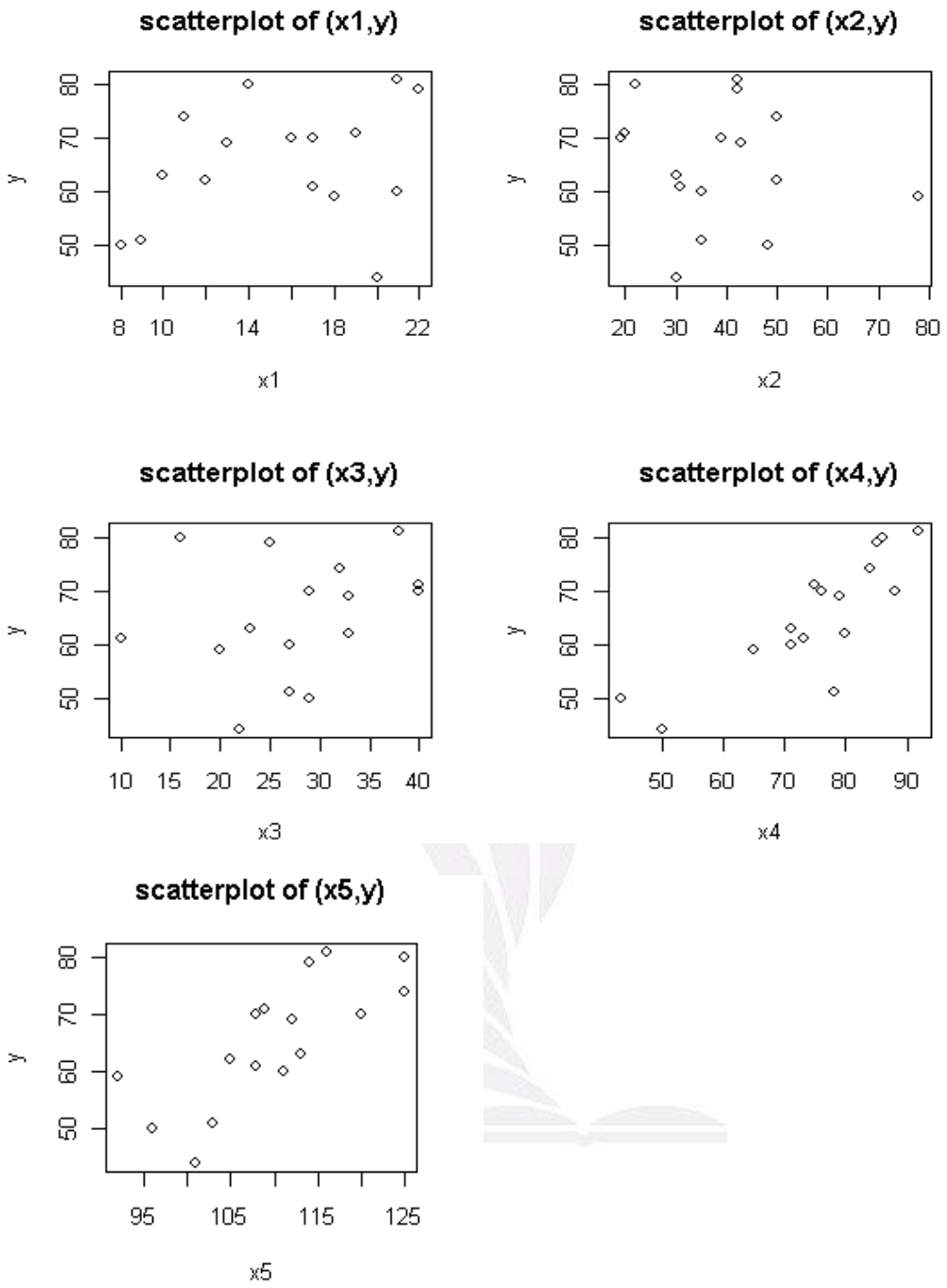


圖 2-4.6 x_i與 y 的散佈圖

由圖 2-4.1～2-4.5 SAS 的 output 可以看出解釋變數 x1、x2、x3 的增減不受反應變數 y 影響，而解釋變數 x4 與 x5 則有隨著反應變數 y 的增加而增加的趨勢，由此推估知，解釋變數 x4 及 x5 與反應變數 y 存在正線性相關。而圖 2-4.6 R 的 output 亦可得到與 SAS 相同的分析結果。

2. 報表主觀的比較：

R 在進行散佈圖的繪製時，可以將多張圖濃縮放置成一大張，在變數間的比較及對應上，相對於 SAS 分開一大張一大張的圖表來的方便美觀並且省空間，因此在散佈圖的繪製上，我們偏愛以 R 軟體進行分析。

第五節　檢測變數間有無多重共線性

一、何謂多重共線性

多重共線性是指在迴歸模式中，某些自變數或所有自變數之間有高度線性相關的現象，此高度相關性會影響配適迴歸線的準確度，因此再配適迴歸前，須先針對解釋變數進行多重共線性的檢測，偵測是否有高度相關的變數存在，若有則需進行轉換加以改善。我們可利用變異數膨脹因素法 Variance Inflation Factor (VIF) 偵測變數間是否存在多重共線性。VIF 主要在量測迴歸係數之變異數，相對於預測變數間的無線性關係之膨脹量，它經常被用來做為多重共線性嚴重程度之指標，當 VIF 超過 10 的情形下，將被視為多重共線性會過度的影響最小平方估計的一項指標訊息。其中 VIF = $\frac{1}{1-R_j^2}$，其中 R_j^2 為以 x_j 為反應變數，剩餘解是變數對於 x_j 的解釋能力。

二、SAS 與 R 的多重共線性檢測之程式碼

1. SAS 程式碼：

```
proc reg;
model y=x1 - x5 /vif;          /*vif: variance inflaction factor */
run;
```

2. R 程式碼：

```
mod1=lm(y~x1+x2+x3+x4+x5)
library(HH)     ## 求 VIF 需要用此套裝程式 ##
v=vif(mod1)
v               ## 列出變數的 v 值 ##
v>10            ## 找 v 值大於 10 的變數##
```

三、SAS 與 R 報表的分析及比較

1. 報表分析：

The REG Procedure
Model: MODEL1
Dependent Variable: Y 閱讀測驗成績

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1473.48271	294.69654	8.73	0.0020
Error	10	337.51729	33.75173		
Corrected Total	15	1811.00000			

Root MSE	5.80962	R-Square	0.8136
Dependent Mean	65.25000	Adj R-Sq	0.7204
Coeff Var	8.90364		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-39.20676	22.66620	-1.73	0.1144	0
x1	單字成績	1	0.43009	0.33400	1.29	0.2268	1.04442
x2	片語成績	1	0.11303	0.11283	1.00	0.3401	1.19931
x3	文法成績	1	0.08268	0.18441	0.45	0.6635	1.06703
x4	對閱讀的期望成績	1	0.41892	0.15766	2.66	0.0240	1.91107
x5	智力測驗成績	1	0.54466	0.23249	2.34	0.0411	2.08148

表 2-5.1　偵測多重共線性 SAS output

```
> mod1=lm(y~x1+x2+x3+x4+x5)
> library(HH)
> v=vif(mod1)   ## 需要用library(HH)##
> v
      x1       x2       x3       x4       x5
1.044416 1.199314 1.067034 1.911074 2.081477
> v>10
   x1    x2    x3    x4    x5
FALSE FALSE FALSE FALSE FALSE
```

表 2-5.2　偵測多重共線性 R output

由表 2-4.1 SAS 的報表可以得知，全模型的參數估計值與模型的解釋能力……等等訊息，並且可得到各個變數的 VIF 值。由表 2-4.1 的 Variance Inflation 得知所有的 VIF 值皆小於 10，表示變數間沒有高度線性相關，不存在多重共線性。

由表 2-4.2 R 的報表可得到比 SAS 報表更精確的 VIF 值(小數點下多一位)，並且加入簡單的一個指令(V>10)，即可判斷出變數間是否具有多重共線性，得到的結果與 SAS 相同，變數間皆沒有高度相關性，不存在多重共線性。

2. 報表主觀的比較：

在偵測多重共線性下，R 和 SAS 的指令都很簡單，但我們認為 R 的 output 相對於 SAS 的簡單明瞭，馬上可以知道各個變數的 VIF 值，而 SAS 則須自行去對應找尋 VIF 值為何，並且加以判斷是否大於 10(具有多重共線性)。因此在此迴歸步驟下我們建議使用 R 軟體進行分析。

第六節　離群值

一、離群值有三大類:

1.殘差（residual）

- 各細格實際觀察人數減去期望人數，又稱為 Δ（delta）值

  $\Delta_{ij} = n_{ij} - \hat{\mu}_{ij}$
- 殘差越大，各細格分佈越不如期望般的出現，兩個變項有某種關聯，殘差越小，表示各細格分佈越接近期望，兩變項無關聯
  - 正殘差值表該細格的觀察次數高於兩個變項無關時的期望值
  - 負殘差值表該細格的觀察次數低於兩個變項無關時的期望值

2.標準化殘差（standardized residual）

- 殘差為未標準化統計量數。將殘差除以標準誤，得到標準化殘差

  $\Delta'_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}}$
- 將標準化殘差平方後加總，即得 Pearson χ2，也就是我們常用的卡方值
- Δ′分佈呈標準化常態分配 N（0,1），可利用常態化 Z 分配進行統計決策

3.調整後標準化殘差（adjusted standardized residual）

- 標準化殘差會隨著邊際期望值的大小變動而產生波動
- 若將標準化殘差以各邊際比率進行調整，得到調整後標準化殘差，可以排除各邊際次數不相等所造成的比較問題

  $adj\Delta' = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1-P_{i.})(1-P_{.j})}}$

二、SAS 與 R 的離群值之程式碼

1. SAS 程式碼：

```
proc reg;
model time=case distance/cli clm p r;
output out=residual p=pred r=resid;
run;
```

2. R 程式碼：

```
s=summary(resid(mod1))
sd(resid(mod1))
var(resid(mod1))
mean(resid(mod1))
std=(resid(mod1)-mean(resid(mod1)))/var(resid(mod1))
boxplot(std)
std<(-3)
```

三、SAS 與 R 報表的分析及比較

1. 報表分析：

Output Statistics

Obs	Residual	Obs	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D
1	-0.7023	1	5.077	-0.138		0.002
2	-1.8794	2	5.458	-0.344		0.004
3	-12.6881	3	5.261	-2.412	****	0.384
4	6.0266	4	5.571	1.082	**	0.027
5	7.0124	5	5.451	1.286	**	0.064
6	4.7804	6	5.204	0.919	*	0.063
7	2.5656	7	5.003	0.513	*	0.028
8	-5.4717	8	4.783	-1.144	**	0.196
9	6.4641	9	4.664	1.386	**	0.336
10	-2.5706	10	5.567	-0.462		0.005
11	-3.9755	11	5.524	-0.720	*	0.015
12	0.7690	12	5.554	0.138		0.000
13	-0.4202	13	5.183	-0.0811		0.001
14	-2.4835	14	4.965	-0.500	*	0.029
15	6.1162	15	4.259	1.436	**	0.569
16	-3.5429	16	5.313	-0.667	*	0.026

Sum of Residuals	0
Sum of Squared Residuals	430.97244
Predicted Residual SS (PRESS)	725.10879

表 2-6.1 離群值 SAS output

```
> s=summary(resid(mod1))
> #sd(resid(mod1))
> #var(resid(mod1))
> #mean(resid(mod1))
> std=(resid(mod1)-mean(resid(mod1)))/var(resid(mod1))
> boxplot(std)
> std<(-3)
    1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16
FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

表 2-6.2 離群值 R output

由表 2-6.1 及表 2-6.2 可以知道，透過 SAS 與 R 進行離群值的分析，兩種軟體的三種的殘差比較之結果相同，皆沒有離群值存在。

2. 報表主觀的比較：

在離群值的比較下，雖然 R 的程式碼比 SAS 較複雜了點，但是因為 R 可以直接輸入判定法則，直接可以了解到哪些點是離群值，而不需像 SAS 利用三項互相比較的方式，因此在離群值的分析比較方法下，我們較推薦 R 軟體進行分析。

第三章　多元迴歸分析法

第一節　選擇重要變數

從眾多的自變數中，利用向前選取法、向後選取法、逐步選取法、全部子集迴歸法有系統的選擇提供較多訊息的重要變數配適迴歸模型。

一、向前選取法(Forward Selection)

一次考慮一個自變數，判斷其貢獻是否已達設定的標準，若是則將其「納入」複(多元)迴歸方程式中。利用向前選取法選取重要變數的步驟如下：首先進入方程式的自變項是與依變項關係最密切者，即與依變項有最大正相關或最大負相關者。接著選取與依變項間的淨相關為最大之自變項，進入迴歸模式。當進入變數之標準化迴歸係數 F 值的機率小於或等於我們可以接受的顯著水準，此變項才可以進入迴歸模式中，且此分析法在選取過程中並不剔除任何已在迴歸方程式中的自變數。

1. SAS 與 R 的向前選取法之程式碼

(1) SAS 程式碼：

```
proc reg;
model y=x1 x2 x3 x4 x5 /selection=forward slentry=0.15;
run;
```

(2) R 程式碼：

```
step(lm(y~1),
     scope = list(upper =~x1+x2+x3+x4+x5+1, lower =  ~1),
     direction=c('forward'))
```

2. SAS 與 R 報表的分析及比較

(1)報表分析：

Forward Selection: Step 1

Variable x4 Entered: R-Square = 0.6810 and C(p) = 5.1149

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1233.34143	1233.34143	29.89	<.0001
Error	14	577.65857	41.26133		
Corrected Total	15	1811.00000			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	13.71715	9.56153	84.92119	2.06	0.1734
x4	0.68940	0.12610	1233.34143	29.89	<.0001

Bounds on condition number: 1, 1

Forward Selection: Step 2

Variable x5 Entered: R-Square = 0.7620 and C(p) = 2.7689

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1380.02756	690.01378	20.81	<.0001
Error	13	430.97244	33.15173		
Corrected Total	15	1811.00000			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-19.94688	18.15428	40.02195	1.21	0.2918
x4	0.47545	0.15206	324.12277	9.78	0.0080
x5	0.45194	0.21485	146.68614	4.42	0.0555

Bounds on condition number: 1.8098, 7.2393

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

Summary of Forward Selection

Step	Variable Entered	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x4	對閱讀的期望成績	1	0.6810	0.6810	5.1149	29.89	<.0001
2	x5	智力測驗成績	2	0.0810	0.7620	2.7689	4.42	0.0555

表 3-1.11 向前選取法 SAS output

```
Start:  AIC=77.66
y ~ 1

       Df Sum of Sq     RSS     AIC
+ x4    1   1233.34  577.66  61.382
+ x5    1   1055.90  755.10  65.668
<none>              1811.00  77.665
+ x1    1    151.78 1659.22  78.264
+ x3    1    134.21 1676.79  78.433
+ x2    1     38.20 1772.80  79.324
Step:  AIC=61.38
y ~ x4

       Df Sum of Sq     RSS     AIC
+ x5    1   146.686 430.97  58.695
<none>              577.66  61.382
+ x1    1    37.651 540.01  62.304
+ x3    1    15.100 562.56  62.959
+ x2    1     1.265 576.39  63.347
Step:  AIC=58.7
y ~ x4 + x5

       Df Sum of Sq     RSS     AIC
<none>              430.97  58.695
+ x1    1    49.348 381.62  58.750
+ x2    1    30.930 400.04  59.504
+ x3    1     9.641 421.33  60.333

Call:
lm(formula = y ~ x4 + x5)

Coefficients:
(Intercept)           x4           x5
   -19.9469       0.4754       0.4519
```

表 3-1.12 向前選取法 R output

由表 3-1.1 SAS 報表可以得知，在顯著水準為 0.15 下，首先挑選變數 x4 進模型，接著將變數 x5 挑入模型中，直到在顯著水準α =0.15 下沒有變數可以再被選進模型中，最後得到兩重要變數為 x_4與x_5。在這兩個解釋變數構成的迴歸模型下，具有 0.762 的解釋能力。

由表 3-1.2 可以得知，利用 R 進行向前選取法選出的重要變數和 SAS 相同，都先後選進了解釋變數 x4 與 x5。且由表 3-1.2 可以得知，R 軟體在向前選取法提供了 AIC 做比較。

(2)報表主觀的比較：

在向前選取法選取重要變數下，SAS 提供了較多的資訊，如：選擇變數 x4 下配適的模型具有的解釋能力為 0.681、加入新變數 x5 額外的解釋能力，以及其 Cp 值。而 R 只提供 AIC 加以判斷，沒有模型的解釋能力……等等其他訊息。因 SAS 在此統計分析下，供應的資訊相對於 R 來的充分許多，因此我們偏愛以 SAS 進行向前選取法選取重要變數。

二、向後消去法(Backward Selection)

所謂的向後消去法，就是將所有的變數納入迴歸模式中，再逐一對變數無法達到設定標準的變數移除，直到所有變數達到標準為止，若是已從迴歸方程式中剔除者，則不再考慮範圍內。

1. SAS 與 R 的向後消去法之程式碼

(1) SAS 程式碼：

```
proc reg;
model y=x1 x2 x3 x4 x5 /selection=backward slstay=0.15;
run;
```

(2) R 程式碼：

```
step(lm(y~x1+x2+x3+x4+x5),direction=c('backward'))
```

2. SAS 與 R 報表的分析及比較

(1)報表分析：

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.8136 and C(p) = 6.0000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1473.48271	294.69654	8.73	0.0020
Error	10	337.51729	33.75173		
Corrected Total	15	1811.00000			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-39.20676	22.66620	100.98584	2.99	0.1144
x1	0.43009	0.33400	55.96691	1.66	0.2268
x2	0.11303	0.11283	33.87233	1.00	0.3401
x3	0.08268	0.18441	6.78429	0.20	0.6635
x4	0.41892	0.15766	238.29868	7.06	0.0240
x5	0.54466	0.23249	185.24039	5.49	0.0411

Bounds on condition number: 2.0815, 36.517

Backward Elimination: Step 1

Variable x3 Removed: R-Square = 0.8099 and C(p) = 4.2010

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1466.69842	366.67460	11.71	0.0006
Error	11	344.30158	31.30014		
Corrected Total	15	1811.00000			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-38.79117	21.80923	99.02190	3.16	0.1029
x1	0.42921	0.32163	55.74118	1.78	0.2090
x2	0.11806	0.10812	37.32256	1.19	0.2982
x4	0.42720	0.15078	251.25516	8.03	0.0163
x5	0.55449	0.22289	193.71460	6.19	0.0302

Bounds on condition number: 2.0629, 24.719

Backward Elimination: Step 2

Variable x2 Removed: R-Square = 0.7893 and C(p) = 3.3068

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1429.37585	476.45862	14.98	0.0002
Error	12	381.62415	31.80201		
Corrected Total	15	1811.00000			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-25.71679	18.37428	62.29704	1.96	0.1869
x1	0.40270	0.32327	49.34829	1.55	0.2367
x4	0.44137	0.15142	270.19756	8.50	0.0130
x5	0.47083	0.21098	158.38360	4.98	0.0455

Bounds on condition number: 1.8709, 14.186

Backward Elimination: Step 3

Variable x1 Removed: R-Square = 0.7620 and C(p) = 2.7689

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1380.02756	690.01378	20.81	<.0001
Error	13	430.97244	33.15173		
Corrected Total	15	1811.00000			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-19.94688	18.15428	40.02195	1.21	0.2918
x4	0.47545	0.15206	324.12277	9.78	0.0080
x5	0.45194	0.21485	146.68614	4.42	0.0555

Bounds on condition number: 1.8098, 7.2393

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

Summary of Backward Elimination

Step	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x3	文法成績	4	0.0037	0.8099	4.2010	0.20	0.6635
2	x2	片語成績	3	0.0206	0.7893	3.3068	1.19	0.2982
3	x1	單字成績	2	0.0272	0.7620	2.7689	1.55	0.2367

表 3-1.21 向後消去法 SAS output

```
Start:  AIC=60.78
y ~ x1 + x2 + x3 + x4 + x5

       Df Sum of Sq     RSS     AIC
- x3    1      6.784 344.30 59.103
- x2    1     33.872 371.39 60.315
<none>                337.52 60.784
- x1    1     55.967 393.48 61.239
- x5    1    185.240 522.76 65.784
- x4    1    238.299 575.82 67.331
Step:  AIC=59.1
y ~ x1 + x2 + x4 + x5

       Df Sum of Sq     RSS     AIC
- x2    1     37.323 381.62 58.750
<none>                344.30 59.103
- x1    1     55.741 400.04 59.504
- x5    1    193.715 538.02 64.245
- x4    1    251.255 595.56 65.871
Step:  AIC=58.75
y ~ x1 + x4 + x5

       Df Sum of Sq     RSS     AIC
- x1    1     49.348 430.97 58.695
<none>                381.62 58.750
- x5    1    158.384 540.01 62.304
- x4    1    270.198 651.82 65.315

       Df Sum of Sq     RSS     AIC
<none>                430.97 58.695
- x5    1     146.69 577.66 61.382
- x4    1     324.12 755.10 65.668
Call:
lm(formula = y ~ x4 + x5)

Coefficients:
(Intercept)           x4           x5
   -19.9469       0.4754       0.4519
```

表 3-1.22 向後消去法 R output

由表 3-2.1 SAS 報表可以得知，在顯著水準為 0.15 下，首先將所有變數納入模型，接著將變數 x3 剔除模型中，再著又將變數 x2 剔除模型中，直到在顯著水準α =0.15 下沒有變數可以再被剔除模型中，最後得到兩重要變數為 x4 與 x5。在這兩個解釋變數構成的迴歸模型下，具有 0.762 的解釋能力。

利用 R 進行向前選取法選出的重要變數，得到表 3-2.2 的結果和 SAS 相同，都先後將 x3、x2 及 x1 踢除出模型中。且由表 3-2.2 可以得知，R 軟體在向後消去法只提供了 AIC 做比較。

(2)報表主觀的比較：

在向後消去法選取重要變數下，SAS 提供了較多的資訊，如：剔除變數 x3 下配適的模型具有的解釋能力為 0.81、各變數額外的解釋能力為何，以及其 Cp 值。而 R 只提供 AIC 加以判斷，沒有模型的解釋能力……等等其他訊息。因 SAS 在此統計分析下，供應的資訊相對於 R 來的充分許多，因此我們偏愛以 SAS 進行向前選取法選取重要變數。

三、逐步迴歸法(Stepwise regression methods)

逐步迴歸選取法是結合「向前選取法」與「向後選取法」而成。預測變數選取過程中輪流以向前、向後選取法選取變數，直到沒有預測變數可以再選進來，亦無預測變數會被去除，這種方式就稱為「逐步迴歸選取法」。

1. SAS 與 R 的逐步迴歸法之程式碼

(1) SAS 程式碼：

```
proc reg;
model y=x1 x2 x3 x4 x5 /selection=stepwise;
run;
```

(2) R 程式碼：

```
step(lm(y~1),
     scope = list(upper = ~x1+x2+x3+x4+x5, lower = ~1),direction=c('both'))
```

2. SAS 與 R 報表的分析及比較

(1)報表分析：

Stepwise Selection: Step 1

Variable x4 Entered: R-Square = 0.6810 and C(p) = 5.1149

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1233.34143	1233.34143	29.89	<.0001
Error	14	577.65857	41.26133		
Corrected Total	15	1811.00000			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	13.71715	9.56153	84.92119	2.06	0.1734
x4	0.68940	0.12610	1233.34143	29.89	<.0001

Bounds on condition number: 1, 1

Stepwise Selection: Step 2

Variable x5 Entered: R-Square = 0.7620 and C(p) = 2.7689

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1380.02756	690.01378	20.81	<.0001
Error	13	430.97244	33.15173		
Corrected Total	15	1811.00000			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-19.94688	18.15428	40.02195	1.21	0.2918
x4	0.47545	0.15206	324.12277	9.78	0.0080
x5	0.45194	0.21485	146.68614	4.42	0.0555

Bounds on condition number: 1.8098, 7.2393

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	x4		對閱讀的期望成績	1	0.6810	0.6810	5.1149	29.89	<.0001
2	x5		智力測驗成績	2	0.0810	0.7620	2.7689	4.42	0.0555

表 3-1.31 逐步迴歸法 SAS output

```
Start:  AIC=77.66
y ~ 1

       Df Sum of Sq     RSS     AIC
+ x4    1   1233.34  577.66  61.382
+ x5    1   1055.90  755.10  65.668
<none>              1811.00  77.665
+ x1    1    151.78 1659.22  78.264
+ x3    1    134.21 1676.79  78.433
+ x2    1     38.20 1772.80  79.324

Step:  AIC=61.38
y ~ x4

       Df Sum of Sq     RSS     AIC
+ x5    1    146.69  430.97  58.695
<none>              577.66  61.382
+ x1    1     37.65  540.01  62.304
+ x3    1     15.10  562.56  62.959
+ x2    1      1.26  576.39  63.347
- x4    1   1233.34 1811.00  77.665

Step:  AIC=58.7
y ~ x4 + x5

       Df Sum of Sq     RSS     AIC
<none>              430.97  58.695
+ x1    1     49.35  381.62  58.750
+ x2    1     30.93  400.04  59.504
+ x3    1      9.64  421.33  60.333
- x5    1    146.69  577.66  61.382
- x4    1    324.12  755.10  65.668

Call:
lm(formula = y ~ x4 + x5)

Coefficients:
(Intercept)           x4           x5
   -19.9469       0.4754       0.4519
```

表 3-1.32 逐步迴歸法 R output

由表 3-3.1 SAS 報表可以得知，在顯著水準為 0.15 下，逐步選取法利用輪流以向前、向後選取法選取變數，挑選變數 x4 進模型，接著將變數 x5 挑入模型中，直到在顯著水準α =0.15 下沒有變數可以再進入模型也沒有變數須被剔除，最後得到兩重要變數為 x4 與 x5。在這兩個解釋變數構成的迴歸模型下，具有 0.762 的解釋能力。

利用 R 進行向前選取法選出的重要變數，得到表 3-3.2 的結果和 SAS 相同。且由表 3-3.2 可以得知，R 軟體在逐步選取法下只提供了 AIC 做比較。

(2)報表主觀的比較：

在逐步選取法選取重要變數下，SAS 提供了較多的資訊，如：挑選出變數 x4 下配適的模型具有的解釋能力為 0.6810、加入新變數 x5 額外的解釋能力，以及其 Cp 值。而 R 只提供 AIC 加以判斷，沒有模型的解釋能力……等等其他訊息。因

SAS 在此統計分析下，供應的資訊相對於 R 來的充分許多，因此我們偏愛以 SAS 進行逐步選取法選取重要變數。

四、全部子集迴歸法(All-subsets Regression)

全部子集的挑選方法有R_p^2、$R_{a,p}^2$、Mallows' C_p、AIC_p與SBC_p五種準則。其中當R_p^2與$R_{a,p}^2$越大，解釋能力越佳；C_p越小且$C_p - p$也越小其模型越佳；AIC_p與SBC_p準則皆是越小越好。

1. SAS 與 R 的全部子集迴歸法之程式碼

(1) SAS 程式碼：

```
proc reg;
model y=x1 x2 x3 x4 x5/selection=adjrsq cp aic sbc best=5;
run;
```

SAS 的 (Cp , p)輔助判斷圖程式碼：

```
proc reg;
model y= x1 x2 x3 x4 x5/selection=cp best=6;
plot cp.*np.
/chocking=green    cmallows=blue
vaxis=0 to 8 by 0.5 haxis=0 to 8 by 0.5 crame=ligr;
symbol1 v=dot c=red;
run;
```

(2) R 程式碼：

```
library(leaps)
x=cbind(x1,x2,x3,x4,x5)
leaps(x,y)
```

R 的 (Cp , p)輔助判斷圖程式碼：

```
library(wle)
mod21=lm(y~x4+x5)
result<-wle.cp(mod21)
plot(result,num.max=7)
```

2. SAS 與 R 報表的分析及比較

(1)報表分析：

Adjusted R-Square Selection Method

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
4	0.7407	0.8099	4.2010	59.1029	62.96581	x1 x2 x4 x5
3	0.7366	0.7893	3.3068	58.7496	61.83991	x1 x4 x5
2	0.7254	0.7620	2.7689	58.6953	61.01305	x4 x5
3	0.7239	0.7791	3.8525	59.5037	62.59408	x2 x4 x5
5	0.7204	0.8136	6.0000	60.7844	65.41998	x1 x2 x3 x4 x5

表 3-1.41 全部子集迴歸法 SAS output

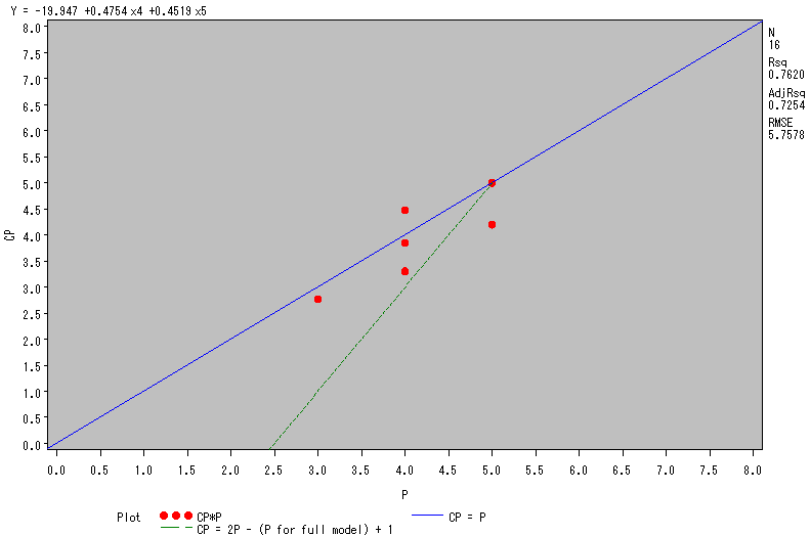


圖 3-1.41 Cp-p 圖的 SAS output

```
$which
      1     2     3     4     5
1 FALSE FALSE FALSE  TRUE FALSE
1 FALSE FALSE FALSE FALSE  TRUE
1  TRUE FALSE FALSE FALSE FALSE
1 FALSE FALSE  TRUE FALSE FALSE
1 FALSE  TRUE FALSE FALSE FALSE
2 FALSE FALSE FALSE  TRUE  TRUE
2  TRUE FALSE FALSE  TRUE FALSE
2 FALSE FALSE  TRUE  TRUE FALSE
2 FALSE  TRUE FALSE  TRUE FALSE
2  TRUE FALSE FALSE FALSE  TRUE
2 FALSE  TRUE FALSE FALSE  TRUE
2 FALSE FALSE  TRUE FALSE  TRUE
2  TRUE FALSE  TRUE FALSE FALSE
2  TRUE  TRUE FALSE FALSE FALSE
2 FALSE  TRUE  TRUE FALSE FALSE
3  TRUE FALSE FALSE  TRUE  TRUE
3 FALSE  TRUE FALSE  TRUE  TRUE
3 FALSE FALSE  TRUE  TRUE  TRUE
3  TRUE FALSE  TRUE  TRUE FALSE
3  TRUE  TRUE FALSE  TRUE FALSE
3 FALSE  TRUE  TRUE  TRUE FALSE
3  TRUE  TRUE FALSE FALSE  TRUE
3  TRUE FALSE  TRUE FALSE  TRUE
3 FALSE  TRUE  TRUE FALSE  TRUE
3  TRUE  TRUE  TRUE FALSE FALSE
4  TRUE  TRUE FALSE  TRUE  TRUE
4  TRUE FALSE  TRUE  TRUE  TRUE
4 FALSE  TRUE  TRUE  TRUE  TRUE
4  TRUE  TRUE  TRUE  TRUE FALSE
4  TRUE  TRUE  TRUE FALSE  TRUE
5  TRUE  TRUE  TRUE  TRUE  TRUE

$label
[1] "(Intercept)" "1"           "2"           "3"           "4"
[6] "5"

$size
 [1] 2 2 2 2 2 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 6

$Cp
 [1]   5.114933 10.372045 37.159694 37.680104 40.524848  2.768899  5.999410
 [8]   6.667549  7.077459  9.312246 10.973871 11.529887 35.369158 38.361675
[15] 38.449831  3.306803  3.852512  4.483254  7.526121  7.940403  8.646283
[22]   9.645221 10.500405 12.342900 36.482953  4.201006  5.003573  5.658194
[29]   9.488323 11.060340  6.000000
```

表 3-1.42 全部子集迴歸法 R output

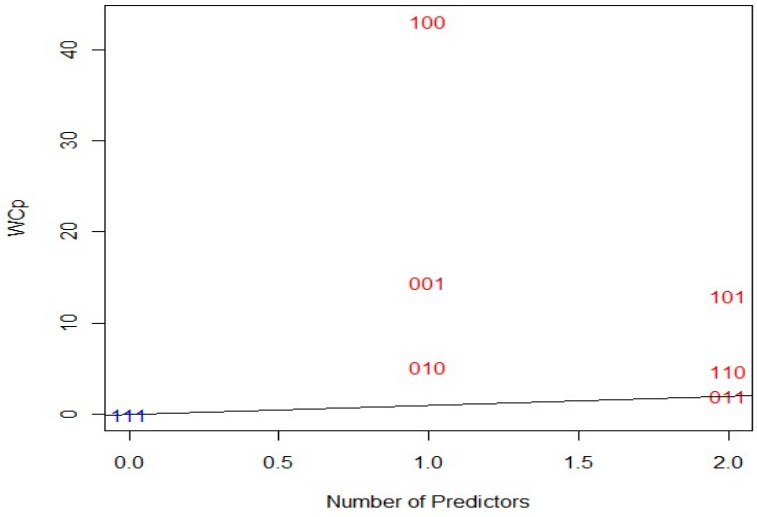


圖 3-1.42 Cp-p 圖的 R output

由表 3-1.41 SAS 報表可知，在全部子集迴歸法選取重要變數下，在只有 x4 與 x5 兩個解釋變數下得到的模型，具有最小的 AIC 與 SBC 值，又由上圖(Cp 與 p 構成的圖)可得知在此模型下具有最小的 Cp 值，且 Cp－p 亦小，而在這兩個解釋變數構成的迴歸模型下，也具有 0.762 的解釋能力。

(2)報表主觀的比較：

在全部子集迴歸方法中，我們可以在 SAS 與 R 的報表中發現，SAS 的提供的訊息相對地比 R 的完整並且清楚許多；就專業內容而言，SAS 的報表中分別顯示了 Adjusted R-square、R-square、C_p 和 AIC 值，但在 R 的報表中只顯示出 C_p 值。除此之外，我們認為 SAS 所跑出的 C_p 圖，比 R 所呈現的圖還要來得美觀，所以對於此方法，我們喜愛並且也建議利用 SAS 進行全部子集迴歸的分析。

第二節　部分 F 檢定

一、何謂部分 F 檢定

主要在檢定部分的參數是否為 0。在原有的解釋變數模型下，額外加入其他變數對於反應變數 y 是否有足夠的解釋貢獻，以至於需要將此變數加入模型中，提高模型解釋反應變數的能力。

假設檢定：$\begin{cases} H_0 : \beta_i = \beta_j = \beta_k = \cdots = 0 \\ H_1 : \beta_i \cdot \beta_j \cdot \beta_k \cdots\cdots \text{不全都為 } 0 \end{cases}$

決策規則：$\begin{cases} \text{若 p-value 小於顯著水準} \alpha, \text{表示拒絕} H_0 : \mu = 0 。 \\ \text{若 p-value 大於顯著水準} \alpha, \text{表示不拒絕} H_0 : \mu = 0 。 \end{cases}$

二、SAS 與 R 的部分 F 檢定之程式碼

1. SAS 程式碼：

```
proc reg data=grade;
model y=x1 x2 x3 x4 x5;
test x1=0, x2=0, x3=0;
run;
```

2. R 程式碼：

```
summary(mod1)
anova(mod1)
F.stat=((151.78 + 26.93 + 130.91)/3)/33.75
F.stat
pf(F.stat, 3, 10, lower=F)
## Getting the P-value (with the appropriate d.f. = (3,10))##
```

三、SAS 與 R 報表的分析及比較

1. 報表分析：

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1473.48271	294.69654	8.73	0.0020
Error	10	337.51729	33.75173		
Corrected Total	15	1811.00000			

Root MSE	5.80962	R-Square	0.8136
Dependent Mean	65.25000	Adj R-Sq	0.7204
Coeff Var	8.90364		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-39.20676	22.66620	-1.73	0.1144
x1	單字成績	1	0.43009	0.33400	1.29	0.2268
x2	片語成績	1	0.11303	0.11283	1.00	0.3401
x3	文法成績	1	0.08268	0.18441	0.45	0.6635
x4	對閱讀的期望成績	1	0.41892	0.15766	2.66	0.0240
x5	智力測驗成績	1	0.54466	0.23249	2.34	0.0411

Test 1 Results for Dependent Variable Y

Source	DF	Mean Square	F Value	Pr > F
Numerator	3	31.15171	0.92	0.4647
Denominator	10	33.75173		

表 3-2.1 部分 F 檢定 SAS output

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -39.20676   22.66620  -1.730   0.1144
x1            0.43009    0.33400   1.288   0.2268
x2            0.11303    0.11283   1.002   0.3401
x3            0.08268    0.18441   0.448   0.6635
x4            0.41892    0.15766   2.657   0.0240 *
x5            0.54466    0.23249   2.343   0.0411 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.81 on 10 degrees of freedom
Multiple R-squared: 0.8136,     Adjusted R-squared: 0.7204
F-statistic: 8.731 on 5 and 10 DF,  p-value: 0.002048

 > anova(mod1)
 Analysis of Variance Table

 Response: y
           Df Sum Sq Mean Sq F value    Pr(>F)
 x1         1 151.78  151.78  4.4968 0.0599661 .
 x2         1  26.93   26.93  0.7980 0.3926773
 x3         1 130.91  130.91  3.8787 0.0772153 .
 x4         1 978.62  978.62 28.9946 0.0003080 ***
 x5         1 185.24  185.24  5.4883 0.0411485 *
 Residuals 10 337.52   33.75
 ---
 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> F.stat=((151.78 + 26.93 + 130.91)/3)/33.75
> F.stat
[1] 3.057975
> ## Getting the P-value (with the appropriate d.f. = (3,10))##
> pf(F.stat, 3, 10, lower=F)
[1] 0.07838727
```

表 3-2.2 部分 F 檢定 R output

由表 3-2.1 SAS 的報表可以知道在 x1、x2、x3、x4、x5 五個解釋變數下，配適的迴歸模型具有 0.7204 的解釋能力。配適模型在給定 x4 和 x5 兩解釋變數下，檢定 x1、x2 與 x3 的參數是否皆為 0，由於 Pr > F = 0.4647 大於顯著水準 $\alpha=0.05$，拒絕$H_0: \beta_1=\beta_2=\beta_3=0$，亦即加入 x1、x2 與 x3 三解釋變數，對於解釋反應變數之變異之能力不高，模型配適時可以不考慮這三個變數。

對 R 軟體下多個指令後，得到報表 3-2.1 亦可得到與 SAS 相同的結果：全模型下，配適的迴歸模型具有 72.04%的解釋能力。在給定 x4 和 x5 兩解釋變數下，檢定 x1、x2 與 x3 是否為 0，由於 *Pr*> F*=0.0784 大於顯著水準$\alpha=0.05$，根據檢定規則，拒絕虛無假設$H_0: \beta_1=\beta_2=\beta_3=0$，亦即 X1、X2 與 X3 可以不被入模型加以配適。

2. 報表主觀的比較：

由於在 SAS 只需簡單的一個指令即可清楚完整得到部分 F 檢定的結果，而 R 卻須自行再加入計算 F 值及其 p-value 之值，相對於 SAS 而言麻煩許多，因此我們偏愛並且建議使用 SAS 軟體進行部分 F 檢定的分析。

第三節　偵測影響點

一、影響點的偵測

影響點即指資料中的觀測點對於迴歸模式的影響，遠超過其他觀測點者。由以下五種方法可以檢定出影響點。

檢定方法	判斷準則
The hat matrix elements h_{ii}	$h_{ii} > 2 \times \frac{p}{n}$
Cook's distance statistic D_i	$D_i > 1$
DFFITS	小樣本：$\lvert DFFITS_i \rvert > 1$ 大樣本：$\lvert DFFITS_i \rvert > 2\sqrt{\frac{p}{n}}$
COVRATIO	$COVRATIO_i > 1 + 3 \times \frac{p}{n}$ $COVRATIO_i < 1 - 3 \times \frac{p}{n}$
DFBETAS	小樣本：$\lvert DFBETAS_i \rvert > 1$ 大樣本：$\lvert DFBETAS_i \rvert > \frac{2}{\sqrt{n}}$
其中 p 為參數個數；n 為樣本數。	

二、SAS 與 R 的部分 F 檢定之程式碼

1. SAS 程式碼：

```
proc reg;
model y=x4 x5 /all;
model y=x4 x5 /Influence;
run;
```

2. R 程式碼：

```
mod21=lm（y~x4+x5）
influence.measures（mod21）
```

三、SAS 與 R 報表的分析及比較

1. 報表分析：

Output Statistics

Obs	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D
1	5.077	-0.138		0.002
2	5.458	-0.344		0.004
3	5.261	-2.412	****	0.384
4	5.571	1.082	**	0.027
5	5.451	1.286	**	0.064
6	5.204	0.919	*	0.063
7	5.003	0.513	*	0.028
8	4.783	-1.144	**	0.196
9	4.664	1.386	**	0.336
10	5.567	-0.462		0.005
11	5.524	-0.720	*	0.015
12	5.554	0.138		0.000
13	5.183	-0.0811		0.001
14	4.965	-0.500	*	0.029
15	4.259	1.436	**	0.569
16	5.313	-0.667	*	0.026

Sum of Residuals	0
Sum of Squared Residuals	430.97244
Predicted Residual SS (PRESS)	725.10879

Output Statistics

Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS Intercept	DFBETAS x4	DFBETAS x5
1	-0.7023	-0.1330	0.2226	1.6282	-0.0712	-0.0252	-0.0598	0.0459
2	-1.8794	-0.3323	0.1014	1.3765	-0.1116	0.0376	0.0621	-0.0641
3	-12.6881	-3.1172	0.1651	0.2569	-1.3864	-0.9105	-0.8783	1.0711
4	6.0266	1.0896	0.0639	1.0234	0.2847	0.0566	0.0320	-0.0417
5	7.0124	1.3231	0.1037	0.9428	0.4501	-0.0697	0.2344	-0.0379
6	4.7804	0.9126	0.1830	1.2724	0.4319	-0.0713	0.3056	-0.0769
7	2.5656	0.4977	0.2450	1.5838	0.2835	-0.2322	-0.0461	0.2095
8	-5.4717	-1.1589	0.3098	1.3400	-0.7764	-0.2127	0.6027	-0.1479
9	6.4641	1.4425	0.3439	1.1995	1.0444	0.9707	0.3361	-0.8812
10	-2.5706	-0.4473	0.0652	1.2943	-0.1181	-0.0325	-0.0003	0.0181
11	-3.9755	-0.7057	0.0796	1.2228	-0.2076	0.0304	0.0935	-0.0796
12	0.7690	0.1331	0.0695	1.3603	0.0364	-0.0031	0.0082	0.0006
13	-0.4202	-0.0779	0.1897	1.5666	-0.0377	0.0264	0.0190	-0.0308
14	-2.4835	-0.4852	0.2563	1.6127	-0.2848	0.2353	0.0750	-0.2256
15	6.1162	1.5041	0.4528	1.3839	1.3683	0.5528	-1.0007	0.0877
16	-3.5429	-0.6519	0.1486	1.3452	-0.2723	-0.1512	-0.1839	0.1941

表 3-3.1 偵測影響點 SAS output

```
> mod21=lm(y~x4+x5)
> influence.measures(mod21)
Influence measures of
         lm(formula = y ~ x4 + x5) :

     dfb.1_    dfb.x4    dfb.x5   dffit cov.r   cook.d    hat inf
1  -0.02516 -0.059847  0.045865 -0.0712 1.628 0.001827 0.2226
2   0.03755  0.062065 -0.064105 -0.1116 1.376 0.004457 0.1014
3  -0.91055 -0.878268  1.071089 -1.3864 0.257 0.383519 0.1651   *
4   0.05655  0.032032 -0.041744  0.2847 1.023 0.026629 0.0639
5  -0.06972  0.234400 -0.037936  0.4501 0.943 0.063841 0.1037
6  -0.07130  0.305578 -0.076906  0.4319 1.272 0.062979 0.1830
7  -0.23217 -0.046078  0.209462  0.2835 1.584 0.028441 0.2450
8  -0.21265  0.602724 -0.147926 -0.7764 1.340 0.195769 0.3098
9   0.97073  0.336053 -0.881163  1.0444 1.199 0.335712 0.3439
10 -0.03250 -0.000271  0.018068 -0.1181 1.294 0.004958 0.0652
11  0.03040  0.093502 -0.079608 -0.2076 1.223 0.014939 0.0796
12 -0.00314  0.008169  0.000582  0.0364 1.360 0.000477 0.0695
13  0.02642  0.019017 -0.030790 -0.0377 1.567 0.000513 0.1897
14  0.23528  0.074976 -0.225614 -0.2848 1.613 0.028732 0.2563
15  0.55281 -1.000733  0.087698  1.3683 1.384 0.568864 0.4528   *
16 -0.15118 -0.183930  0.194051 -0.2723 1.345 0.025863 0.1486
```

表 3-3.2 偵測影響點 R output

由表 3-3.1 SAS 報表可以得知，在有 x4 與 x5 兩個解釋變數的模型中，參數個數 p＝3，樣本數 n＝16 屬於小樣本。故當 Cook's distance statistic $D_i > 1$；

$|DFFITS_i| > 1$；$|DFBETAS_i| > 1$；The hat matrix elements $h_{ii} > 2 \times (3/16) =$ 0.375 及 $COVRATIO_i > 1 + 3 \times 3/16) = 1.5625$ or $COVRATIO_i < 1 - 3 \times (3/16) =$ 0.4375 時，此觀察值及有可能是影響點。由報表可知，第 3 筆與第 15 筆觀測值有可能是影響點。

由表 3-3.2 R 報表亦可得到與 SAS 相同的結果，且 R 軟體將偵測影響點的各種檢定方法中判斷準則的數值算出，即可從星號(＊)看出，此觀察值及有可能是影響點。由報表可知，第 3 筆與第 15 筆觀測值有可能是影響點。

2. 報表主觀的比較：

在偵測影響點下，SAS 需要先計算出各判斷準則的值，在由我們自行判斷比較，而 R 報表會直接用星號(＊)告知影響點，不需要再加以計算。雖然 R 可以簡單看出影響點，但卻不能得知觀察值在何種檢定方法下，被視為影響點。即使如此，我們依舊喜愛且建議使用 R 報表來偵測影響點，因為對於樣本數只有十六個的我們，當然是輕而易舉，但樣本數增加時，要將大樣本中的影響點判斷出來顯得比較麻煩， 使用 R 的報表來偵測影響點，不僅省時又可以快速偵測出影響

第四章 殘差診斷分析

在迴歸模型中，若誤差項ε_i為獨立之常態隨機變數且平均數為0，變異數為常數 σ^2，則殘差 e_i 會反映出誤差項ε_i的性質，這就是殘差分析的基本假設，也是一種用來檢驗統計模型之適當性的有效方法，因此我們對於誤差項ε_i的四種性質分別加以檢定。誤差項 ε_i 的假設：

(1) $E(\varepsilon_i) = 0$。

(2) $Var(\varepsilon_i) = \sigma^2$。

(3) $Cov(\varepsilon_i, \varepsilon_j) = 0$：$\forall i \neq j$。

(4) ε_i服從常態。

第一節 檢測殘差平均是否為零

一、檢測殘差平均是否為零

假設檢定：$\begin{cases} H_0 : \mu = 0 \\ H_1 : \mu \neq 0 \end{cases}$

決策規則：$\begin{cases} \text{若 p-value 小於顯著水準}\alpha\text{，表示拒絕}H_0 : \mu = 0\text{。} \\ \text{若 p-value 大於顯著水準}\alpha\text{，表示不拒絕}H_0 : \mu = 0\text{。} \end{cases}$

二、SAS 與 R 檢定殘差平均是否為零之程式碼

1. SAS 程式碼：

```
proc univariate normal plot ;
var student;
run;
```

2. R 程式碼：

```
mod3=lm(y~x4+x5)
student3=(resid(mod3)-mean(resid(mod3)))/var(resid(mod3))
library(stats)
t.test(student3)
```

三、SAS 與 R 報表的分析及比較

1. 報表分析：

Moments

N	16	Sum Weights	16
Mean	0.01826663	Sum Observations	0.29226606
Std Deviation	1.05275934	Variance	1.10830223
Skewness	-0.4859285	Kurtosis	0.30449805
Uncorrected SS	16.6298722	Corrected SS	16.6245334
Coeff Variation	5763.2931	Std Error Mean	0.26318984

Basic Statistical Measures

Location		Variability	
Mean	0.01827	Std Deviation	1.05276
Median	-0.10971	Variance	1.10830
Mode	.	Range	3.84781
		Interquartile Range	1.58367

Tests for Location: Mu0=0

Test		-Statistic-	-----p Value------	
Student's t	t	0.069405	Pr > |t|	0.9456
Sign	M	-1	Pr >= |M|	0.8036
Signed Rank	S	5	Pr >= |S|	0.8209

Tests for Normality

Test		--Statistic---	-----p Value------	
Shapiro-Wilk	W	0.941263	Pr < W	0.3648
Kolmogorov-Smirnov	D	0.116654	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.043798	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.333823	Pr > A-Sq	>0.2500

```
                 The UNIVARIATE Procedure
       Variable:  student   (Studentized Residual)

                  Extreme Observations

   ------Lowest------            ------Highest-----

         Value      Obs               Value      Obs

   -2.411775          3           0.918521         6
   -1.143876          8           1.081824         4
   -0.719711         11           1.286447         5
   -0.666848         16           1.386054         9
   -0.500156         14           1.436032        15

 Stem Leaf                     #             Boxplot
    1 1344                     4             +-----+
    0 59                       2             |     |
    0 1                        1             |  +  |
   -0 311                      3             *-----*
   -0 7755                     4             +-----+
   -1 1                        1                |
   -1                                           |
   -2 4                        1                |
      ----+----+----+----+

                      Normal Probability Plot
     1.25+                                   * *****    *
         |                               ******
         |                           +++++
         |                       ***** *
         |                 ******
         |             +++++
         |        +++++
    -2.25+   ++++ *
          +----+----+----+----+----+----+----+----+----+----+
               -2        -1         0        +1        +2
```

表 4-1.1　檢測殘差平均是否為零 SAS output

```
        One Sample t-test

data:  student3
t = 0, df = 15, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.09941134  0.09941134
sample estimates:
  mean of x
4.98733e-18
```

表 4-1.2　檢測殘差平均是否為零 R output

由表 4-1.1 SAS 的報表可得知有基本的統計量，還有檢定假設殘差平均是否為零的 p-value 值，這些 p-value 皆大於顯著水準α=0.05，故不拒絕H_0：$\mu = 0$，意即殘差平均為 0。亦可看出殘差的 normal probability plot 呈一直線，符合常態分配，且對於誤差是否服從常態分配的假設之 p-value 也都大於顯著水準α=0.05，故不拒絕 H_0，意即誤差服從常態分配。

由 2-1.2 R 的報表亦可得到相同的結果，因 p-value=1，大於顯著水準α=0.05，故不拒絕H_0：$\mu = 0$，意即殘差平均為 0。

2. 報表主觀的比較：

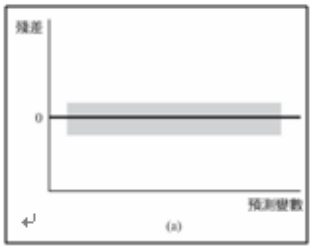
SAS 的程式碼只需短短打出幾行指令，不但可以進行殘差平均是否為 0 的檢定，還可得知其基本統計量，並且得到殘差是否符合常態的假設檢定之結果，因此在檢測殘差的基本假設上，我們喜愛使用 SAS 來進行分析。

第二節 檢測殘差變異數是否為常數

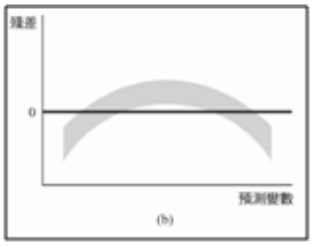
一、檢測殘差變異數是否為常數

利用殘差圖判斷變異數是否為常數，若出如圖(a)，即表示變異數為常數，圖(b)或圖(c)即表示異數不為常數：

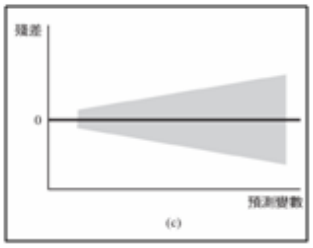
(a)變異數為常數



(b)變異數為不規則



(c)變異數隨變數X遞增



二、SAS 與 R 的殘差變異數是否為常數之程式碼

1. SAS 程式碼：

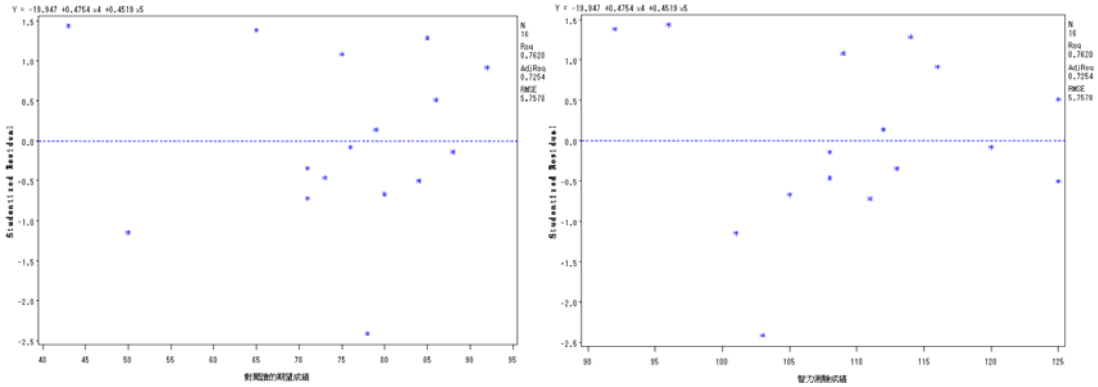
```
PROC REG;
MODEL y=x4 x5 ;
plot(student. rstudent.)*(x4 x5);
symbol1 v=star c=brown;
run;
```

2. R 程式碼：

```
mod3=lm(y~x4+x5)
par(mfrow=c(2,2))
yhat=fitted(mod3)    ##predicted values##
plot(yhat,student3)
plot(x4,student3)
plot(x5,student3)
```

三、SAS 與 R 報表的分析及比較

1. 報表分析：



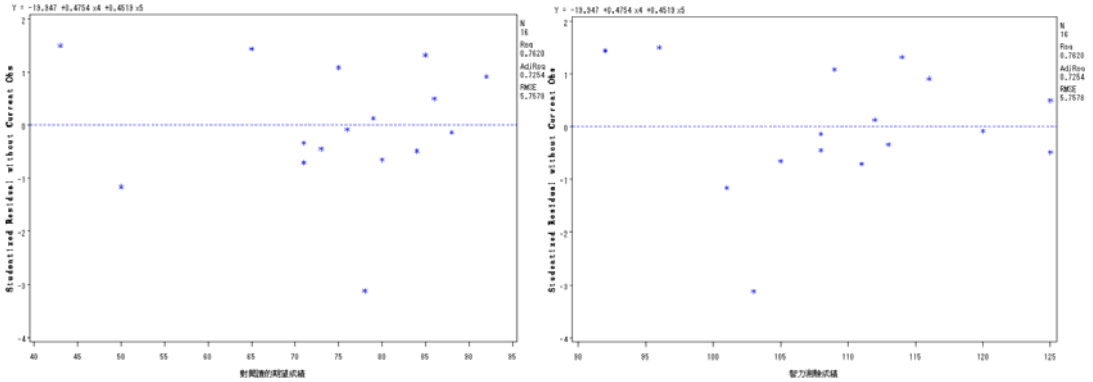


圖 4-2.1　檢測殘差變異數是否為常數 SAS output

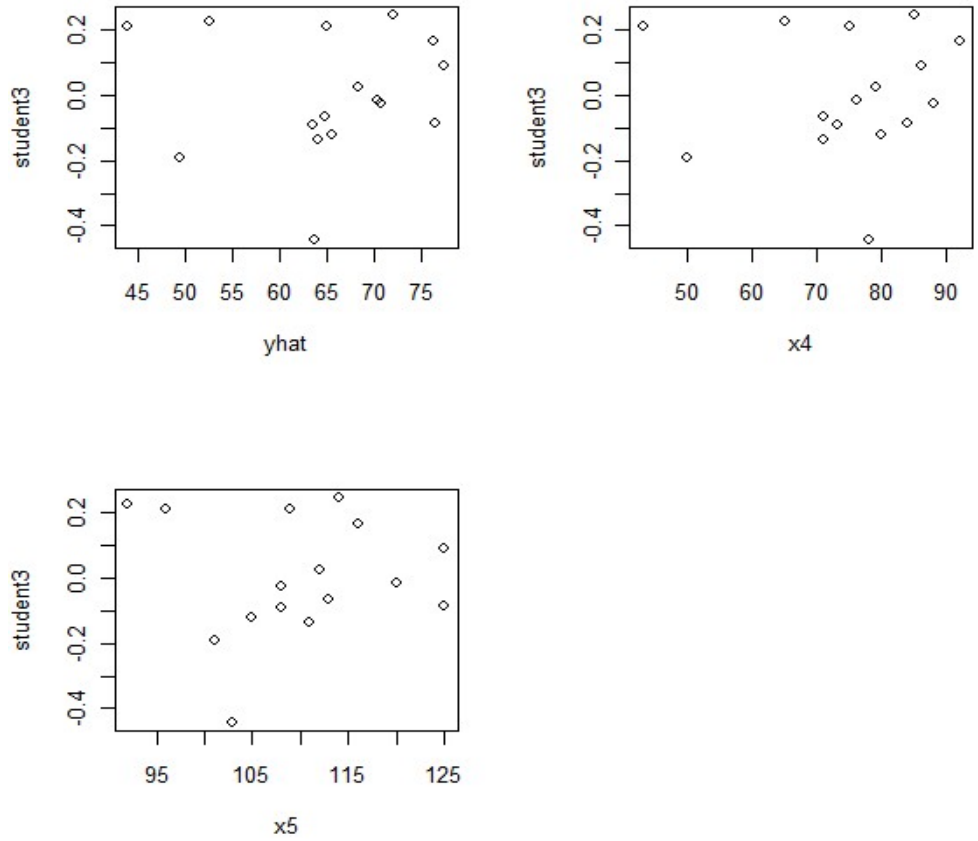


圖 4-2.2　檢測殘差變異數是否為常數 R output

由圖4-2.1和圖4-2.2的殘差散佈圖可看出，解釋變數的殘差和標準化殘差的散佈圖，沒有明顯不規則、遞增或遞減的情況，因此判斷殘差變異為常數，符合誤差項的基本假設Var(ε_i)＝σ^2。

2. 報表主觀的比較：

SAS 的 output 是一大張一大張分開的圖表，如需看出差異則要統一整理起來比較，相對於 R 加上指令 par(mfrow=c(2,2))來的麻煩。而兩個軟體在圖表上的顏色以及數據分佈的標示點，都可另外下指令以不同形式表現出來。因此在檢測殘差變異數是否為常數上，我們喜愛使用 R 來進行分析。

第三節　檢測殘差殘差是否為獨立

一、檢測殘差是否相互獨立

當殘差不存在自我相關性時，才是一個好的配適模型，我們利用自我相關的 Durbin-Watson 檢定判定第一階自我迴歸模型的自我相關參數 ρ 是否為零，若是則 ε_t 相互獨立。

DW 檢定統計量檢測規則：

當 DW 值為 2 時，表示此模型的殘差不具自我相關性。

若 DW 值介於 0 到 2 之間則表示殘差存在正自我相關。

而 DW 值介於 2 到 4 之間則表示殘差存在負自我相關。

假設檢定：$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho > 0 \end{cases}$　$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho < 0 \end{cases}$

檢定規則：

當 Pr<DW 的值小於顯著水準 α=0.05 時，表示顯著存在正自我相關。

當 Pr>DW 的值小於顯著水準 α=0.05 時，表示顯著存在負自我相關。

二、SAS 與 R 的殘差相互獨立檢定之程式碼

1. SAS 程式碼：

```
proc autoreg;
model y=x4 x5/dwprob;
run;
```

2. R 程式碼：

```
library(lmtest)
dwtest(y~x4+x5,alternative="greater")
dwtest(y~x4+x5,alternative="less")
```

三、SAS 與 R 報表的分析及比較

1. 報表分析：

The AUTOREG Procedure

Dependent Variable Y 閱讀測驗成績

Ordinary Least Squares Estimates

SSE	430.972436	DFE	13
MSE	33.15173	Root MSE	5.75775
SBC	106.419086	AIC	104.10132
Regress R-Square	0.7620	Total R-Square	0.7620
Durbin-Watson	2.2401	Pr < DW	0.6784
Pr > DW	0.3216		

NOTE: Pr<DW is the p-value for testing positive autocorrelation, and Pr>DW is the p-value for testing negative autocorrelation.

Variable	DF	Estimate	Standard Error	t Value	Approx Pr > |t|	Variable Label
Intercept	1	-19.9469	18.1543	-1.10	0.2918	
x4	1	0.4754	0.1521	3.13	0.0080	對閱讀的期望成績
x5	1	0.4519	0.2149	2.10	0.0555	智力測驗成績

表 4-3.1　檢測殘差是否相互獨立 SAS output

```
> #dwtest(lmtest)
> library(lmtest)
Loading required package: zoo
> dwtest(y~x4+x5,alternative="greater")

        Durbin-Watson test

data:  y ~ x4 + x5
DW = 2.2401, p-value = 0.6784
alternative hypothesis: true autocorrelation is greater than 0

> dwtest(y~x4+x5,alternative="less")

        Durbin-Watson test

data:  y ~ x4 + x5
DW = 2.2401, p-value = 0.3216
alternative hypothesis: true autocorrelation is less than 0
```

表 4-3.2　檢測殘差是否相互獨立 R output

由表 4-3.1 SAS 的報表得知 DW=2.2401，且 Pr>DW 為 0.3216 及 Pr<DW 為 0.6786，兩個值皆大於顯著水準α=0.05，不拒絕$H_0: \rho = 0$，表示殘差不具有相關性，故ε_i相互獨立，符合誤差項的基本假設 Cov(ε_i, ε_j)＝0；$\forall\, i \neq j$。

由表 4-3.2 R 的報表得知 DW=2.2401，正負相關檢定的 p-value 值也皆大於顯著水準α=0.05，不拒絕$H_0: \rho = 0$，表示殘差不具有相關性，故ε_i相互獨立，亦符合誤差項的基本假設 Cov(ε_i, ε_j)＝0；$\forall\, i \neq j$。

2. 報表主觀的比較：

檢測殘差有無自我相關時，利用 SAS 軟體只需一個指令就可以得到正自我相關、負自我相關的檢定結果及 DW 值，可直接由同一張報表看出檢定是否符合假

設；而 R 報表必須分別下指令看出各別的正負相關檢定是否符合，相對之下，SAS 的程式碼較為簡單，檢定也可從同一張報表看出。因此在檢測殘差是否相互獨立上，我們喜愛使用 SAS 來進行分析。

第四節　檢測誤差是否為常態

一、檢測殘差是否為常態

我們利用Shapiro-Wilk、Kolmogorov-Smirnov、Cramer-von Mises、Lillie與Anderson-Darling 這些方法來做誤差像是否為常態的假設檢定。

假設檢定：$\begin{cases} H_0：誤差服從常態 \\ H_1：誤差不服從常態 \end{cases}$

檢定規則：若 p-value 小於顯著水準α，表示拒絕 H_0：誤差服從常態。

若 p-value 大於顯著水準α，表示不拒絕 H_0：誤差服從常態。

二、SAS 與 R 的殘差是否為常態檢定之程式碼

1. SAS 程式碼：

```
proc univariate normal plot ;
var student;
run;
```

2. R 程式碼：

```
library(nortest)
ad.test(student3)
cvm.test(student3)
lillie.test(student3)
pearson.test(student3)
sf.test(student3)
```

```
qqnorm(student3,main="Normal Q-Q plot",xlab="Theoretical Quantiles",
        ylab="Sample Quantiles",plot.it=TRUE,datax=FALSE,col='red')
qqline(student3,datax=FALSE)
```

三、SAS 與 R 報表的分析及比較

1. 報表分析：

Basic Statistical Measures

Location		Variability	
Mean	0.01827	Std Deviation	1.05276
Median	-0.10971	Variance	1.10830
Mode	.	Range	3.84781
		Interquartile Range	1.58367

Basic Statistical Measures

Location		Variability	
Mean	0.01827	Std Deviation	1.05276
Median	-0.10971	Variance	1.10830
Mode	.	Range	3.84781
		Interquartile Range	1.58367

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	0.069405	Pr > |t|	0.9456
Sign	M	-1	Pr >= |M|	0.8036
Signed Rank	S	5	Pr >= |S|	0.8209

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.941263	Pr < W	0.3648
Kolmogorov-Smirnov	D	0.116654	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.043798	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.333823	Pr > A-Sq	>0.2500

The UNIVARIATE Procedure
Variable: student (Studentized Residual)

Extreme Observations

------Lowest------		------Highest-----	
Value	Obs	Value	Obs
-2.411775	3	0.918521	6
-1.143876	8	1.081824	4
-0.719711	11	1.286447	5
-0.666848	16	1.386054	9
-0.500156	14	1.436032	15

Stem	Leaf	#	Boxplot
1	1344	4	+-----+
0	59	2	| |
0	1	1	| + |
-0	311	3	*-----*
-0	7755	4	+-----+
-1	1	1	|
-1			|
-2	4	1	|

----+----+----+----+

Normal Probability Plot

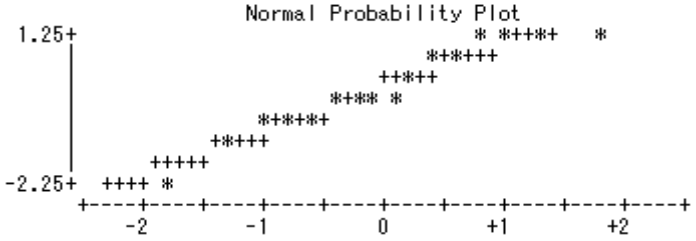


表 4-4.1　檢測殘差是否為常態 SAS output

```
> library(nortest)
> ad.test(student3)

        Anderson-Darling normality test

data:  student3
A = 0.3868, p-value = 0.3465

> cvm.test(student3)

        Cramer-von Mises normality test

data:  student3
W = 0.0481, p-value = 0.5159

> lillie.test(student3)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  student3
D = 0.1263, p-value = 0.7093

> pearson.test(student3)

        Pearson chi-square normality test

data:  student3
P = 4.125, p-value = 0.3894

> sf.test(student3)

        Shapiro-Francia normality test

data:  student3
W = 0.9306, p-value = 0.2134
```

表 4-4.2　檢測殘差是否為常態 R output

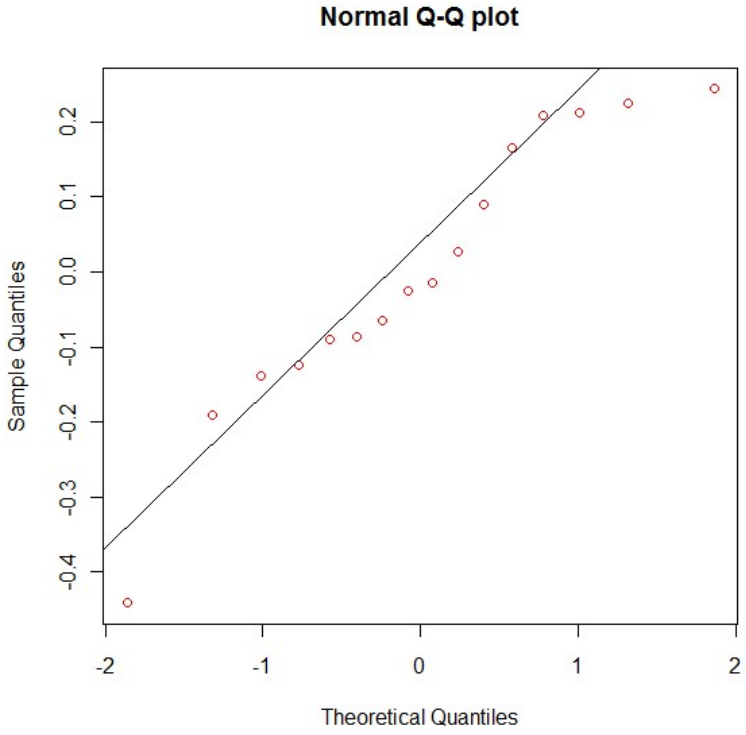


圖 4-4.1　檢測殘差是否為常態 Q-Q plot 圖

由表 4-1.1 SAS 的報表可得知有基本的統計量，還有檢定假設殘差平均是否為零的 p-value 值，這些 p-value 皆大於顯著水準α=0.05，故不拒絕H_0：$\mu = 0$，意即殘差平均為 0。亦可看出殘差的 normal probability plot 成一直線，符合常態分配，且對於誤差是否服從常態分配的假設之 p-value 也都大於顯著水準α=0.05，故不拒絕 H_0，意即誤差服從常態分配。

由表 4-4.2 R 報表可得知所有方法的檢定法結果之 p-value 皆大於顯著水準α=0.05，不拒絕H_0：誤差服從常態，表示誤差服從常態分配。而從圖 4-4.3 可看出 normal Q-Q plot 大致上呈一直線，符合常態分配。

2. 報表主觀的比較：

SAS 的程式碼只需短短打出幾行，就可看出各種檢定以及另外可得知出基本統計量，而從常態機率圖，也很清楚的看出是否符合常態；相對 R 程式碼須個別打出檢定的指令，所以不能只從一個 output 就看出是否符合假設。因此在檢測殘差是否為常態上，我們喜愛使用 SAS 來進行分析。

第五節　最終模型

根據基本統計量分別找出資料的平均數及標準差，並加以分析解釋；接著使用相關係數分析變數間的相關程度高或是低；然後藉由散佈圖概略的看出變數間的基本關係，是正相關、負相關或是無相關；再用多重共線性的檢測方法，檢測解釋變數間是否存在高度相關性；重要的是選出重要的變數，並配適迴歸模型；最後對配適出的模型進行殘差基本假設的檢定，分別有殘差平均是否為零、判斷殘差變異數是否為常數、檢定殘差是否相互獨立、檢定誤差是否為常態……等，在所有假設均無誤下，得到我們的最終終模型如下：

$$\hat{Y}_i = -19.9469 + 0.4754X_4 + 0.4519X_5$$

第五章 分析結果總結表

統計分析方法	統計套裝軟體 R	統計套裝軟體 SAS
基本統計量		☺
相關係數		☺
散佈圖	☺	
多重共線性	☺	
離群值	☺	
選擇重要變數		☺
部分 F 檢定		☺

偵測影響點	☺	
殘差分析法		☺
費用	☺	

表 5-1　SAS 與 R 的 output 優劣比較表

第六章　結論與建議

由表 5-1，可以發現依據分析方法的不同，SAS 與 R 皆有其優劣之處，而每種套裝軟體都有本身它的優缺點，所以才會有如此多種的套裝軟題發明，以補及不足之處。在這次的比較報告中，我們建議可以依照數據的需要、分析方法的不同，將套裝軟體 SAS 跟 R 同時交互使用，以達到最大的效益；相對地，每個分析方法也有其適合的軟體，希望這份報告可以做為未來學弟妹研究的範疇，將統計套裝軟體發揮淋漓盡致，發現它更多的功能，以達到最佳的分析結果。

參考文獻

1. 資料來源分析：
   http://webclass.ncu.edu.tw/~tang0/Chap12/Sas12.htm#範例 12.1
2. SAS 與 R 的程式碼範例：
   http://www.stat.sc.edu/~hitchcock/stat704.html
3. 書籍：
   SAS 1-2-3　作者: 彭昭英 / 出版社: 臺北市　儒林　2010[民 99]
4. 維基百科：
   http://zh.wikipedia.org/zh/SAS%E7%B3%BB%E7%BB%9F